

A Concentration-Dependent Analysis Method for High Density Protein Microarrays

Ovidiu Marina,^{†,||} Melinda A. Biernacki,[‡] Vladimir Brusic,[†] and Catherine J. Wu^{*,†,§}

Cancer Vaccine Center and Division of Hematologic Neoplasia, Dana-Farber Cancer Institute, and Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, and Case Western Reserve University School of Medicine, Cleveland, Ohio 44106

Received December 29, 2007

Protein microarray technology is rapidly growing and has the potential to accelerate the discovery of targets of serum antibody responses in cancer, autoimmunity and infectious disease. Analytical tools for interpreting this high-throughput array data, however, are not well-established. We developed a concentration-dependent analysis (CDA) method which normalizes protein microarray data based on the concentration of spotted probes. We show that this analysis samples a data space that is complementary to other commonly employed analyses, and demonstrate experimental validation of 92% of hits identified by the intersection of CDA with other tools. These data support the use of CDA either as a preprocessing step for a more complete proteomic microarray data analysis or as a stand-alone analysis method.

Keywords: immune responses • proteomic • protein microarray • antigen identification • ProtoArray

Introduction

The rapidly growing technology of protein microarrays has been used for identifying protein-protein interactions,^{1,2} discovering disease biomarkers,^{3,4} and identifying DNA-binding specificity by protein variants.⁵ It has enabled the characterization of humoral immune responses^{6–8} and the determination of antibody specificity.^{9,10} It also promises to revolutionize drug development^{11,12} and the study of cell-level biochemical interactions.² With the advent of commercially available high-density protein microarrays, large-scale proteomic experiments are now feasible, even if limited by the repertoire of protein probes selected by the microarray's manufacturer.¹³ Applications in immunology include the definition of the serologic signatures of and potential antigenic targets for immunization and immune monitoring of various disease states.

Essential to the employment of this proteomic technology is the development of suitable analytical methods. DNA microarray methodologies,^{6,7} including the direct use of existing DNA microarray software suites, such as Express Yourself,¹⁴ SNOMAD¹⁵ and TM4¹⁶, have been proposed or implemented for protein microarray analysis.^{17–19} Methods specifically developed for protein microarrays have also been recently published,^{20,21} but the robustness of these techniques is unclear. None of the proposed methods have gained ac-

ceptance outside the initial publication, and untested methods are commonly employed in published protein microarray studies. Methods for the analysis of antibody microarrays have also been developed,²² and may be applicable to protein microarray analysis. Prospector is a free software provided by the commercial protein microarray manufacturer Invitrogen for the analysis of their ProtoArrays.²³ ProtoArrays have already been applied in a study for autoantibodies in neuromyelitis optica⁴ and an ovarian cancer screen.¹³ No publications to date, however, either detail the methodology implemented by Prospector, or evaluate the reliability of its results. A better understanding of analysis tools for protein microarray data, both descriptive and experimentally validated, is necessary for the interpretation of raw data and the production of reliable and reproducible results across experiments.

The importance of spot concentration, or the amount of protein available for interaction on the microarray, has been previously recognized as an important factor for which correction is needed.^{10,18,20,24,25} Several methods have been proposed for correcting for variation in spot concentration. One common approach is to divide the raw signal by the estimated protein concentration on the array.^{10,20,26} This, however, disproportionately increases the significance of signal of low-concentration spots. Setting cutoffs for which concentrations will be normalized²⁰ helps with this bias, but does not fully model the underlying signal distribution. Prospector software, the most accessible analytical tool and therefore the most likely primary analysis tool for ProtoArray data, does not address the issue of spot concentration. In practice, identified interactions are biased toward high-concentration protein spots on the microarray. To address this concern, we developed a method to correct for protein spot concentration and performed

* To whom correspondence should be addressed. Catherine J. Wu, M.D., Dana-Farber Cancer Institute, Harvard Institutes of Medicine, Room 416, 77 Avenue Louis Pasteur, Boston, MA 02115. Phone: 617-632-5943. Fax: 617-632-3351. E-mail: cwu@partners.org.

[†] Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School.

[‡] Division of Hematologic Neoplasia, Dana-Farber Cancer Institute.

[§] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School.

^{||} Case Western Reserve University School of Medicine.

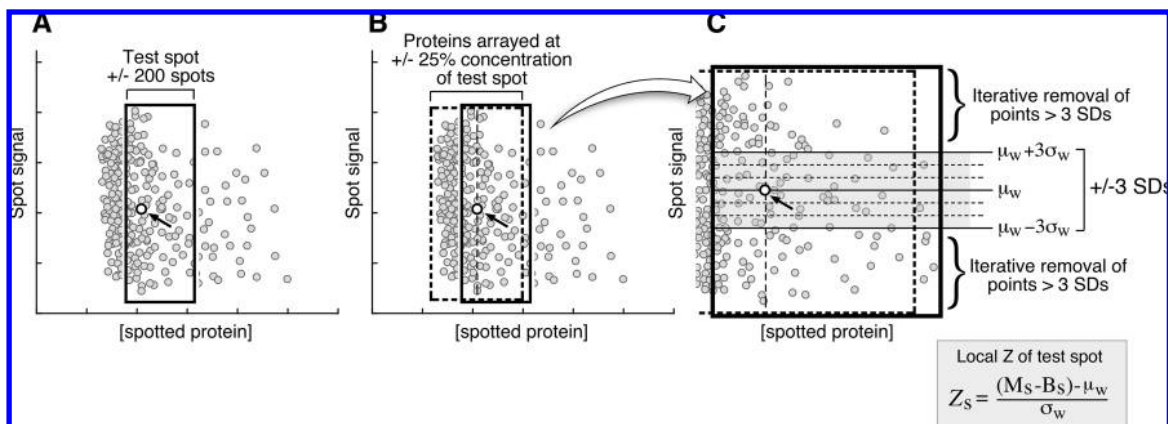


Figure 1. Graphical depiction of the CDA algorithm. (A) A window containing the nearest protein spots by concentration is defined for a given protein spot, which is identified by the arrow. The window contains up to 200 spots above and 200 spots below the given protein spot. (B) The range of neighboring protein spots is then restricted to spots with concentration within 25% of the concentration of the given spot. (C) Protein spots outside of three standard deviations σ_w from the local mean μ_w are iteratively removed from further calculation. Spot signals S_s are calculated as the median fluorescence of the spot (M_s) minus the median local background (B_s). Z-scores are then calculated from these using the local distribution, as shown in eq 5. Although the given protein spot is depicted as having a spot signal equal to the local mean, it may have any value.

separate experimental validation of predicted significant interactions to clarify the effect of correcting for the protein concentration. These transformations were applied to a data set generated from high-density Invitrogen ProtoArrays that were probed with plasma immunoglobulin derived from leukemia patients who responded to immunotherapy. Prospector and our method were used to screen for potential tumor-associated antigens eliciting significant antibody interactions that would merit further detailed downstream characterization. Herein, we have explored the effect of concentration correction on protein microarray data analysis with the aim of improving the identification of significant protein–protein interactions and of increasing the true positive identification of serum antibody interactions while keeping false-positive identifications low.

Materials and Methods

Patient Samples. Heparinized peripheral blood samples were obtained from patients with chronic lymphocytic leukemia (CLL) or chronic myelogenous leukemia (CML). Patients were enrolled on clinical research protocols of donor lymphocyte infusion (DLI) that were approved by the Human Subjects Protection Committee at the Dana-Farber Cancer Institute. Plasma was isolated by removal of the plasma layer after centrifugation of whole blood and cryopreserved at -80°C until the time of analysis.

Protein Microarray Probing. Commercial protein arrays (v3 human ProtoArray, Invitrogen, Carlsbad, CA) were probed with pre- and post-DLI plasma samples from patients with CML and CLL. Arrays were processed at 6°C , according to the manufacturer's instructions. In brief, protein arrays were blocked with 1% bovine serum albumin (BSA)/0.1% Tween 20/1× phosphate buffered saline (PBS) for 1 h. A total of 120 μL of serum diluted 1:150 in probe buffer (1% BSA/0.05% Triton X-100/0.5 mM dithiothreitol (DTT)/5 mM MgCl_2 /5% glycerol/1× PBS) was added and a coverslip was placed above the array surface. After 90 min, the array was washed three times for 8 min each with probe buffer. The array was then incubated in 1:2000 anti-human AlexaFluor 647 conjugate (Invitrogen, Carlsbad CA) in probe buffer for 90 min. The wash was then repeated, before the arrays were dried at room temperature.

Arrays were scanned at $5\ \mu\text{m}$ resolution using a GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA) at 100% power and 600 gain setting. Lot-specific protein spot definitions²³ provided by the manufacturer were manually aligned and fit to the scanned image data. Fluorescence intensities were quantified using GenePix Pro 5.0 (Molecular Devices) using the default settings of local background subtraction.

Data Analysis Using ProtoArray Prospector. Prospector 4.0 (Invitrogen, Carlsbad, CA) is free software provided for use with GenePix Pro result files from processed ProtoArray protein microarrays.²³ Of the three analysis methods included in Prospector, the methods protein–protein interaction (PPI) and immune response profiling (IRP) are both recommended as applicable to serum screening (Prospector 4.0 User Guide).²⁷ The PPI analysis method calculates Z-scores Z_k for each feature using the background-corrected feature signal X_k , array-wide mean μ_s and standard deviation σ_s :

$$Z_k = \frac{X_k - \mu_s}{\sigma_s} \quad (1)$$

For PPI, high signal proteins are defined as the proteins whose average Z-score between the two replicate spots for that protein is greater than the standard value of 3.²⁸ This default threshold can be modified by the user.

In contrast, IRP is based on Chebyshev inequality (CI) *p*-values, which are calculated using the signal distribution of buffer-only control features on the array with mean μ_n and standard deviation σ_n as:

$$\text{CI } p\text{-value} = \begin{cases} 1 & X_k \leq \mu_n + \sigma_n \\ \left(\frac{\sigma_n}{X_k - \mu_n} \right)^2 & X_k > \mu_n + \sigma_n \end{cases} \quad (2)$$

For IRP, high signal proteins are defined as the proteins whose CI *p*-value is less than $1/N_{\text{array}}$, where N_{array} is the number of proteins on the array. This default threshold can also be modified by the user.

Although the computation for PPI and IRP shown in eqs 1 and 2 appear different, the two methods are closely related. Ignoring spots whose CI *p*-value is set to 1, the comparison

Concentration-Dependent Protein Microarray Analysis

between the calculated CI p -value and the default constant threshold for IRP becomes:

$$\left(\frac{\sigma_n}{X_k - \mu_n}\right)^2 < \frac{1}{N_{\text{array}}} \quad (3)$$

Taking the inverse and square root of each side gives:

$$\frac{X_k - \mu_n}{\sigma_n} > \sqrt{N_{\text{array}}} \quad (4)$$

Comparing eqs 1 and 4 shows that PPI and IRP share the same formula with different parameter values for significance determination. Although eq 3 also has a negative solution in addition to eq 4, the negative solution is not applicable because the CI p -value evaluates to 1, or not high-signal, in eq 2 for all X_k to which this would apply. The positive solution, listed in eq 4, appropriately decides all spots whose CI p -value is defaulted to 1 in eq 2 as not high-signal, as can be shown by substituting the cutoff value of $X_k = \mu_n + \sigma_n$. As a result, for both PPI and IRP, a simple numeric cutoff can be calculated that can be directly applied against the raw background-corrected fluorescence measurements from the microarray by solving eqs 1 and 4 for the smallest high-signal X_k .

Comparison of multiple arrays using Prospector reports all proteins with high signal on at least one of the microarrays being compared. For our study, we defined hits, or significant interactions, as the proteins that had high signal post-treatment, but not pretreatment, when comparing pre- and post-treatment microarrays for each patient individually, using PPI and IRP separately. Only comparisons among arrays from a single lot are allowed by Prospector.

Concentration-Dependent Analysis (CDA). The CDA algorithm, shown in Figure 1, corrects for the effect of microarray protein concentration on the measured feature signals by selecting a group of proteins of similar concentration to define the distribution for use as local statistics. For each feature, the spot signal S_s is calculated as the median spot fluorescence signal minus the median local background fluorescence signal. Each spot has a measured protein concentration C_s which is provided by the manufacturer, is lot-specific, and is based on anti-GST probing of a subset of printed microarrays from each lot.²³ The S_s are sorted by C_s , and a window of 200 spots below and 200 spots above C_s is defined and extended to include all spots of concentration equal to the uppermost and lowermost concentration values included in the window. The window size is then reduced, if necessary, so as to not encompass spots below $0.75C_s$ or above $1.25C_s$. S_s outside of 3 standard deviations σ_w from the local mean μ_w are removed iteratively, until the list of spots stops changing. Z_s are calculated for each spot as:

$$Z_s = \frac{S_s - \mu_w}{\sigma_w} \quad (5)$$

Using Z_{pre} from the pretreatment array and Z_{post} from the post-treatment array, two variables, Z_{delta} and Z_{mult} , are defined:

$$Z_{\text{delta}} = Z_{\text{post}} - \max(0, Z_{\text{pre}}) \quad (6)$$

$$Z_{\text{mult}} = \frac{Z_{\text{post}}}{\max(1, Z_{\text{pre}})} \quad (7)$$

To discriminate true signal change from random signal variation, both replicate spots on the array are required to have Z_{delta} and Z_{mult} above a threshold n . We used a stringent cutoff

of $n = 5$ to define a hit, or significant interaction, while the standard cutoff of $n = 3^{28}$ was used to define the CDA gray zone. These criteria for hit selection were selected empirically to yield a number of hits similar to PPI and IRP for the analyzed data. Because proteins can be ordered by calculated significance, the cutoff values can be easily varied. The lower limits enforced for Z_{pre} in eqs 6 and 7 limit the effect small Z_{pre} has on calculated significance. Data analysis for CDA was performed using Excel (Microsoft, Seattle, WA).

Hit Validation by Immunoprecipitation Assay. A subset of serum antibody–protein interactions identified by the PPI, IRP and CDA methods were selected for validation by immunoprecipitation of *de novo* synthesized recombinant protein using the same patient plasma as was used for the microarray screening, with immunoprecipitated antibody–antigen interactions visualized on immunoblot. Plasmids encoding DNA sequences of interest were acquired in a Gateway plasmid vector (Invitrogen, Carlsbad, CA)²⁹ or in other vectors and PCR-cloned into the Gateway storage vector (PlasmID, Cambridge, MA;³⁰ Dana-Farber Cancer Institute, Boston, MA;³¹ American Tissue Culture Collection, Manassas, VA; Open Biosystems, Huntsville, AL; RZPD, Berlin, Germany).³² The DNA sequences were then shuttled into a mammalian expression vector (gift of Wagner Montor, Harvard Institute of Proteomics, Cambridge, MA), and protein was synthesized *in vitro* using Transcend biotinylated lysine tRNA (Promega, Madison, WI) and rabbit reticulocyte lysate (Promega) following the manufacturer's protocol. To immunoprecipitate protein, 5 μL of serum was combined with 5 μL of reticulocyte lysate and product. The mixture was incubated at 4 °C for 1 h. A total of 40 μL of protein A sepharose CL-4B beads (GE Healthcare, U.K.) and 450 μL of 1 $\times$ PBS were then added, and the mixture was incubated at 6 °C with rotation for 1 h. The beads were washed five times by centrifuging at 1000 rpm and 4 °C for 3 min, removing the supernatant and adding 1 mL of 0.05% Tween 20/1 $\times$ PBS. The beads were then centrifuged at 14 000 rpm and 4 °C for 5 min, the supernatant was removed, 25 μL of 3 $\times$ SDS Laemmli sample buffer was added, and the mixture was heated to 90 °C for 8 min. A total of 15 μL of the final product was loaded onto a Ready Gel Tris-Hcl (Bio-Rad, Hercules, CA) and run according to the manufacturer's instructions. Following the transfer of protein onto nitrocellulose membranes (Protran BA85, Whatman, Florham Park, NJ), these membranes were blocked overnight in blocking buffer (0.5% Tween 20/1 $\times$ TBS). Antigen–antibody interactions were detected using a streptavidin–HRP conjugate (MP Biomedicals, Solon, OH) at 1:5000 dilution in blocking buffer followed by Supersignal West Femto detection substrate (Pierce Biotechnology, Rockford, IL). Hits were considered validated if a stronger band was present in the product of immunoprecipitation with post-treatment compared to pretreatment sera.

Results

Concentration as a Significant Correction Factor for Analysis of Protein Microarrays. Donor lymphocyte infusion (DLI) is an established immunotherapy for the treatment of patients with post-transplant relapsed hematologic malignancies, including chronic myelogenous leukemia (CML) and chronic lymphocytic leukemia (CLL). To identify new targets of humoral immunity associated with this immunotherapy, we probed 4 pairs of protein microarrays from two printing lots with pre- and post-DLI plasma from 2 CLL and 2 CML patients, with one lot used for each disease. Another set of 3 microarrays

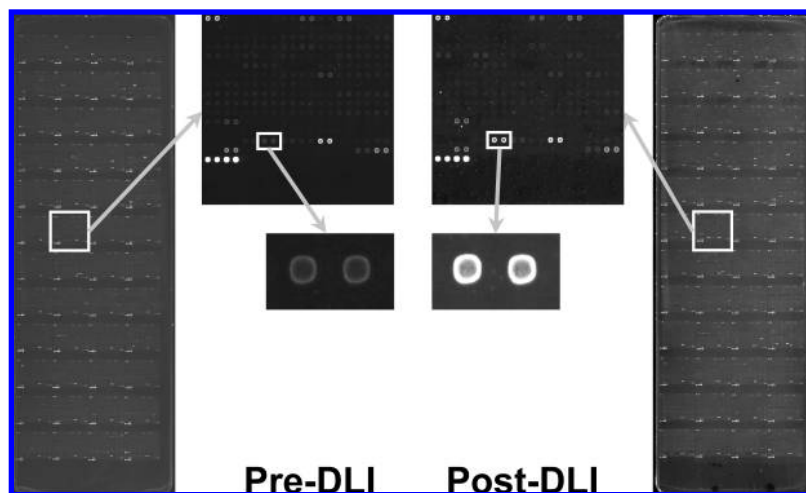


Figure 2. Representative example of a pair of protein microarrays (Protoarray v3, Invitrogen) screened with plasma diluted 1:150 taken pre- and post-donor lymphocyte infusion (DLI) from a CLL patient. Each microarray contains approximately 5000 open reading frames expressed as GST-fusion proteins in an insect cell line that are spotted in duplicate. Antibody–protein binding was detected using an Alexa-Fluor 647 conjugated anti-human IgG antibody. As shown in the inset, while most spots have unchanged reactivity pre- versus post-treatment, some have clearly differential antibody binding.

was probed with pre-DLI and two distinct post-DLI plasma samples from a third CML patient. These studies thus yielded six pre- to post-DLI comparisons, two for the two CLL and four for the three CML patients. As shown in Figure 2, the majority of protein spots yielded equivalent raw fluorescence signal when tested with pre- versus post-treatment plasma for a given patient, with only a subset of proteins showing differential signal that may prove to be of biological significance. Low levels of background noise and minimal artifact were observed for the analyzed arrays.

With more conventional immunoproteomic platforms such as bacteriophage expression libraries, statistical analysis of the strength of observed interactions with respect to characteristics such as synthesized protein amount is not possible. This, however, is feasible with defined-antigen protein microarrays, as spotting concentration of the protein and the exact amino acid sequence of the protein probe is provided. Our initial analysis of the CLL and CML experimental data using the PPI and IRP Prospector algorithms with default cutoff settings resulted in lists of hits that contained a disproportionate number of proteins that were printed at higher concentration when compared to the content of the microarrays. Because proteins that are antigenic in the post-immunotherapy setting for leukemia are not expected to have a correlation with the concentration of protein on the microarray, proteins of interest should be randomly distributed within the concentration range on the microarray. However, as shown in Figure 3, whereas the concentration of 55% of proteins spotted on the microarrays was less than 200 nM, and of 75% was less than 500 nM, only 5–7% and 9–13% of hits identified by PPI and IRP, respectively, had concentrations within these ranges. Moreover, we observed great differences in concentration between spots of the same protein on different printing runs of the commercial microarrays used, as of the roughly 10 000 protein spots on the microarrays of two of the lots used, 46% had less than half and 2% had more than double the concentration of the equivalent spots in one lot compared to the other.

Taken together, these observations highlighted the important contribution of arrayed protein concentration to the analysis of protein–antibody interactions. For these reasons, we de-

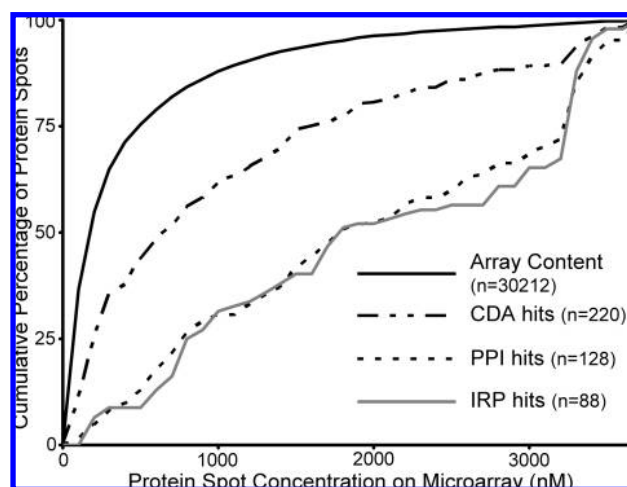


Figure 3. Cumulative percentages of microarray protein and hit spots, as selected by the three methods among the CML and CLL data sets. Interactions identified among immunoglobulin or control protein probes were not included in the plots. The distribution of microarray protein concentrations is shown as the top curve, highlighting that a majority of the proteins on the microarrays have a low protein concentration. If true hits were randomly distributed among the proteins on the microarrays, the cumulative distributions of identified hits by each method should mirror the curve of microarray protein content. As can be seen, however, all three methods identify a disproportionate share of their hits from among proteins of high concentration, and have a bias against low-concentration proteins. PPI and IRP have a strong bias, identifying 40% and 43% of their hits within the top 2.5% of proteins by concentration, while CDA has less of a bias, identifying 14% of its hits within this group.

veloped a novel algorithm to correct for the effect of protein concentration on measured signal. This method, called the concentration dependent algorithm (CDA), calculates the significance of each individual protein signal relative to a window of other arrayed proteins of similar concentration, and is shown in Figure 1. Windows are first defined as a numeric range of spots, then are restricted by concentration and iterative spot removal based on spot signal. The window sizes decrease

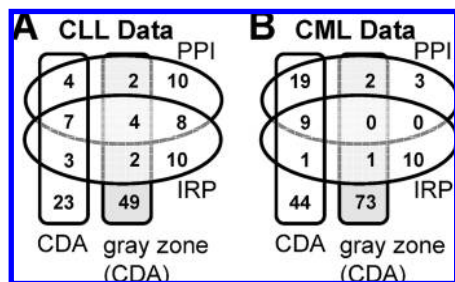


Figure 4. Summary of the identified hits using each of the three methods (CDA, PPI, and IRP) using default cutoffs. Shown are the hits identified for the (A) CLL and (B) CML data sets. The gray zone denotes the additional hits identified by using the lower CDA cutoff threshold. Only a minority of proteins fall within the overlap between PPI or IRP and CDA. The addition of the CDA hits to the PPI and IRP hits nearly doubles the identified number of significant interactions.

after this processing, to an extent that varies from array to array, as spot removal is data-driven. Taking 2 microarrays from the CLL and 2 microarrays from the CML data sets as representative examples, we found that after iterative spot removal 95–96% of spots had 300 or more neighbors left for local distribution and significance calculation, 2–4% had 200–299 neighbors left, and 1–2% had less than 200 neighbors left. Only 30–66 protein spots on any of these microarrays had less than 50 neighbors defining their local signal distribution. These protein spots are found at the extremes of concentration, either nearly 0 or at the high end of reported concentrations. No identified CLL hit or gray zone protein was in this latter group.

Identification of Antibody Targets in Patients Following Immunotherapy. To identify a broad selection of possible antigens eliciting significantly greater antibody reactivity following immunotherapy compared to prior to treatment, we applied the PPI, IRP, and CDA methods to data sets generated from the six pre- to post-treatment comparisons for CLL and CML. As shown in Figure 3, all three pre- to post-comparison methods generate result sets of hits that are biased towards high-concentration proteins. However, we found that this bias was less for CDA than for PPI or IRP, since hits identified from the top 2.5% of proteins by concentration comprised only 14% of CDA hits, compared to 40–43% of PPI and IRP hits.

Application of all three analysis methods with default cutoff settings to our screening studies revealed a total of 73 and 89 unique antibody–protein interactions identified by one or more of these methods as hits for the CLL and CML patients, as summarized in panels A and B, respectively, of Figure 4. For CDA, we defined an intermediate range of potential hits, termed the gray zone, between the hits and unselected proteins, using a less stringent cutoff of $n = 3$. Including the gray zone, the candidate lists were expanded to 122 CLL and 162 CML unique antibody–protein interactions. A selection of positive control proteins, primarily known autoantigens spotted onto the microarray at concentrations 2- to 30-fold higher than the highest-concentration probe protein, were not of interest for our study and were excluded from the analysis.

While the CLL data set yielded roughly equivalent numbers of hits for each analysis method—35 hits by PPI, 34 hits by IRP, and 37 hits by CDA—the CML data set identified more hits by CDA (73), compared to the PPI (33) or IRP (21) methods. Overall, the degree of overlap between proteins identified between CDA and the union of PPI and IRP was a relatively

modest 20% (14 of 73) within the CLL data set and 33% (29 of 89) within the CML data set. Although the numeric overlap of the extended CDA and gray zone list with the union of PPI and IRP is greater, the percent overlap drops to 18% (22 of 122) for CLL and 20% (32 of 162) for CML. An even smaller number of proteins were identified as hits by all three of the methods—only seven of the CLL and nine of the CML hits. Finally, when combining the hits identified in the CML and CLL data sets, 8% (13 of 162) were identified by PPI alone, 12% (20 of 162) by IRP alone, and 41% (67 of 162) by CDA alone. These results show that CDA analysis is complementary to the PPI and IRP methods, and that CDA analysis nearly doubles, and with the gray zone triples, the number of prevalidation significant interactions identified on the microarrays.

Rates of Nonsignificant Interactions Identified as Significant Vary by Analysis Method. Thirty of the 5000 proteins found on the microarrays are immunoglobulin (Ig) sequences, as identified using MatchMiner,³³ and are natural targets of the secondary anti-human IgG antibody used for signal detection on the microarrays. Because the same secondary antibody is used in a similar manner across microarrays, the significance of fluorescent signal would not be expected to differ among microarrays. We found, however, that CDA, PPI, and IRP varied in the extent to which they avoid identifying these presumably nonrelevant interactions as significant. CDA identified 110 hits, or significant interactions, across all CLL and CML comparisons—nearly as many as the 68 hits determined by PPI and 55 hits determined by IRP combined—but 0% of the identified interactions were Ig sequences. Of the interactions identified by PPI and IRP outside of their respective intersections with CDA, 4 of 29 (14%) PPI interactions and 11 of 35 (31%) IRP interactions identified Ig sequences. Both PPI and IRP identified Ig interactions as significant in nearly every comparison. The CDA gray zone is less robust in excluding these interactions, as immunoglobulins comprised 6 of 133 (5%) interactions in this set. Although Ig false positives can be easily removed from data analysis by data mining, the rate of known nonsignificant interaction identification may reflect the inclusion of additional falsely identified interactions within these data sets. Moreover, it supports the idea that the intersection of CDA with PPI and IRP can decrease the rate of false-positive hit identifications while increasing true positive identifications.

Comparison of Analysis Methods. To better characterize the spot signal distribution of the hits identified by each method, we mapped these hits against the cutoff thresholds for each method using the data set generated from CLL patient B as a representative example. As shown in Figures 4 and 5, the default cutoffs for each method select different but overlapping sets of hits. The PPI (Figure 5A) and IRP (Figure 5B) methods define constant cutoffs, shown as the solid lines on each graph, for the determination of significance. These constant cutoffs are calculated using the raw spot signal values from the pre- and post-treatment data sets. Hits, or significant antibody–protein interactions, are less than pretreatment cutoff, depicted by a vertical line, and above the post-treatment cutoff, depicted by a horizontal line. In contrast, the default cutoff function for CDA is defined using transformed data, specifically local Z-scores, and as shown in Figure 5C is nonlinear. Significant antibody–protein interactions have a relatively higher post-versus pretreatment significance, and fall to the left of and above the cutoff line.

We observed that PPI and IRP are closely related methods, sharing a similar formula with different parameter values for

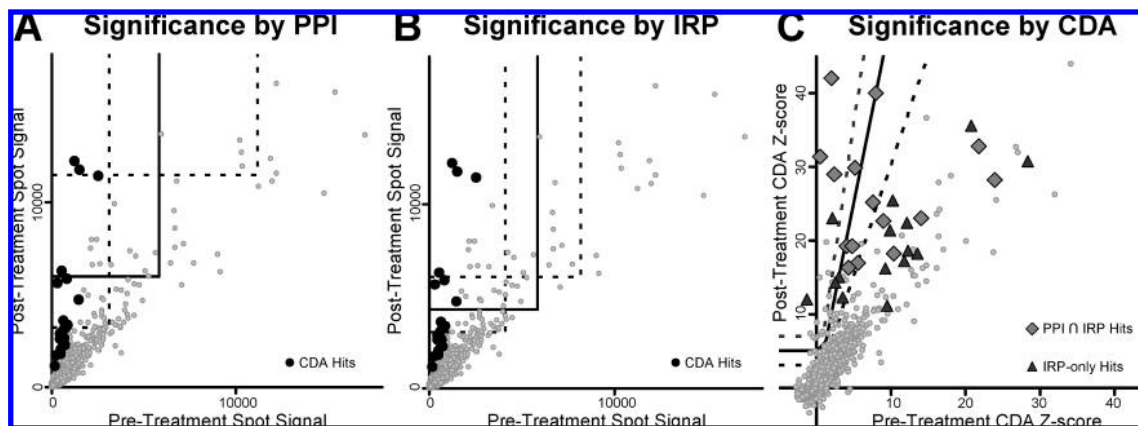


Figure 5. Comparison of the transformations and overlap of hits among the different analysis methods using pre- and post-treatment data for CLL patient B. A single point is plotted for each protein on the microarray as the average signal of the two replicate spots for that protein (A and B), or the lowest significance calculated for that protein across the two replicate spots (C). (A) Overlay of the CDA hits (black circles) on the plot of raw spot signal, with the PPI default cutoff (3, black line) and more stringent (6) and lenient (1.5) PPI cutoffs (dashed lines) shown. (B) Overlay of the CDA hits (black circles) on the plot of raw spot signal, with the IRP default cutoff ($1/N_{\text{array}}$, calculated as its spot signal equivalent, black line) and more stringent ($1/2N_{\text{array}}$) and lenient ($2/N_{\text{array}}$) cutoffs (dashed lines) shown. As can be seen in both (A) and (B), as the cutoff value is varied, a discontinuous list of hits is identified by the PPI and IRP methods. Moreover, CDA identifies a large number of hits with spot signal far below, not just near, the thresholds used by PPI and IRP for significance calculation. (C) Overlay of the PPI (diamond) and IRP (both diamond and triangle) hits on the CDA-calculated significance for the proteins on the microarray. For this patient, all PPI hits were also identified as IRP hits. The black line is the CDA default cutoff of $n = 5$, and the dashed lines denote $n = 10$ (above) and $n = 3$ (the default CDA gray zone cutoff, below). As can be seen, for CDA when the cutoff is increased, a subset of hits is identified, and when the cutoff is decreased, an inclusive superset of hits is identified. Moreover, the significance calculated by CDA for PPI and IRP hits outside of the region of overlap with CDA is quite variable.

significance determination (see Materials and Methods). Because PPI and IRP share the same binary cutoff function and only differ in the constants defining the cutoffs, they both identify a discontinuous list of significant interactions as cutoff constants are varied, shown by the dotted lines in Figure 5A,B. Thus, altering their respective cutoffs results in additional proteins being added to and a subset of the previously identified proteins being removed from the set of hits. On the other hand, CDA defines a continuous list of significant interactions, with hits identified by more stringent cutoffs always being included in sets of hits determined by more lenient cutoffs.

Figure 5 also provides visual representation of the overlap of identified significant interactions among the three methods. The 21 hits identified by CDA for CLL patient B are shown in Figure 5A,B, showing that the majority of identified hits have relatively low raw feature signals compared to the overall microarray data. By comparison, the hits identified by PPI and IRP are shown in Figure 5C, overlaid on the transformed significance space of CDA. Over half of these hits are below the gray zone defined by CDA, and nearly a quarter are estimated to be of nearly equal significance pre- and post-treatment by CDA. These results demonstrate that CDA and the Prospector algorithms sample different and only partially overlapping areas of the data set space.

Experimental Validation of Identified Hits. To determine whether identified hits represent true antibody–protein interactions, we tested the validity of a subset of the CLL microarray results using immunoprecipitation. Interactions with proteins consisting of immunoglobulin sequences, 2 derived from PPI analysis and 3 from IRP, were excluded from the validation list. These interactions were not counted as failed validations as they remained untested. As we were limited by the amount of available banked pre- and post-treatment patient plasma sample, only a subset of the identified interactions was tested.

We acquired DNA sequences of the hits selected for validation and then synthesized recombinant proteins. Significant spot signals can result from true interactions, erroneous measurement, and interactions with improperly synthesized protein or co-precipitated protein complexes. By using sequence-verified template and a different protein synthesis method than employed in manufacture of the microarray, we attempted to avoid the same confounding factors related to protein synthesis and, thus, separate true interactions from screening artifacts. Because true interactions can be due to the recognition by serum antibodies of linear epitopes, conformational epitopes, or their post-translational modifications, we used a mammalian expression system to most accurately replicate *in vivo* protein synthesis. As shown by a representative example in Figure 6A, we then probed each newly synthesized protein for interaction with pre-DLI compared to post-DLI patient serum using immunoprecipitation and visualized the results by immunoblot. Interactions were designated as validated if increased immunoprecipitated protein was visualized using post- compared to pretreatment plasma.

A summary of the validation studies for the CLL data sets is presented in Figure 6B. Overall, 34 of 54 selected hits (63%) were successfully validated, that is, represent true positive interactions. Successful validations were observed for 21 of 29 CDA hits (72%), for 15 of 23 PPI hits (65%), and 14 of 23 IRP hits (61%). Strikingly, successful validation was observed in 12 of 13 (92%) proteins identified in common between CDA and either the PPI or IRP methods. On the other hand, outside of this intersection, validations were successful for only 9 of 16 CDA-only hits (56%), 6 of 13 PPI hits (46%), and 5 of 13 IRP hits (38%). Finally, 2 validated proteins from patient A which are not hits in patient B and 2 validated proteins from patient B which are not hits in patient A were also subjected to validation using sera that had not identified them as hits. As expected, all four interactions failed validation. Taken together,

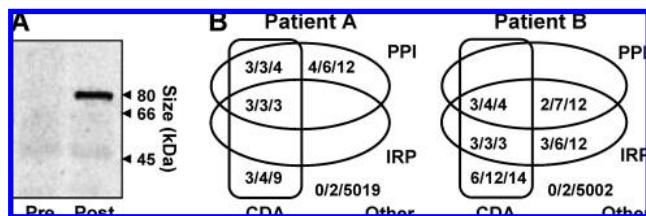


Figure 6. (A) Sample immunoblot of validation for a protein identified as a hit for patient B by CDA but not by either PPI or IRP. (B) Comparison of the validation for CLL proteins identified as hits by the three methods, PPI, IRP, and CDA, for each of the CLL patients. Numbers are shown as validated/tested/total hits in each category. For example, for patient A by CDA alone, 9 hits were identified, 4 were tested, and 3 were successfully validated. Although the validation rates for PPI, IRP, and CDA were similar (65%, 61%, and 72%, respectively), hits identified by the intersection of PPI or IRP and CDA had a 92% validation rate, with only a single identified interaction not replicated experimentally. Therefore, CDA not only provides a method for increasing the sensitivity of significant interaction identification, by identifying additional hits that can be validated, but the intersection of CDA with PPI or IRP can be used to increase the specificity of hit identification, by selecting the hits that are very likely to be successfully validated.

these results demonstrate that the CDA method shows similar performance to the union of PPI and IRP and doubles the number of validated hits relative to each of these methods. Additionally, antigens identified by the intersection of CDA with either of the other methods could be consistently validated experimentally. No consistent differences in the intensity of spot signal or spot concentration were observed between the hits that succeeded as compared to failed validation within the groups of CDA only, PPI and IRP only, or their intersection (Supplemental Tables 1 and 2 in Supporting Information).

Nine of the 57 CLL CDA gray zone hits were also tested, either because they overlapped with PPI or IRP, had identical sequence to tested CDA hits, or were otherwise selected (data not shown). Five of these 9 (56%) interactions, including 2 of the 3 interactions identified only by the CDA gray zone, were validated successfully. These data demonstrate that true interactions are also present among the hits identified using less stringent criteria by CDA.

Effect of CDA Rank Order on Identification of Antigens Found in Common between Analysis Methods. As we had found the highest rate of true positive interactions among antigens identified by the intersection of the three analysis methods, we next analyzed our prevalidation data sets to establish the projected rate of overlap of PPI or IRP with CDA hits, as a function of sorting based on the rank order of the CDA-defined hit list, as shown in Figure 7. Because Prospector provides only a binary assignment for high-signal and therefore for hit determination, there is no *a priori* rationale for sorting hits identified by PPI and IRP, and these are all combined with the top-ranked CDA hit at the beginning of the plot. In Figure 7, prevalidation data sets generated from the two CLL patients are shown as representative examples, consisting of 42 CDA hits from CLL patient A (16 CDA and 26 gray zone) and 52 CDA hits from CLL patient B (21 CDA and 31 gray zone), organized in order of continuous ranked significance by CDA. We observed that lengthening the list of hits identified by CDA was associated with a gradual increase in the number of antigens identified in common between CDA and PPI or IRP, although this rate of intersection eventually reached a plateau, such that

the combination of CDA with PPI or IRP hits represented about half of CDA hits and a third of CDA hits with gray zone entries. Therefore, the addition of CDA to the PPI and IRP methods also provides a means for ranking of ProtoArray hits.

Discussion

Target antigen identification remains a high priority in the study of cancer, infectious disease, and autoimmunity, and the process of antigen identification has been greatly accelerated by the recent availability of protein microarray technology. It is thus imperative that appropriate analysis tools are available, since robust analysis of the data can provide significant cost-savings of labor and resources, which are required for detailed downstream characterization. With DNA microarrays, the majority of differences that arise among the results from various studies arise from differences in the methodology employed for data analysis,³⁴ and protein microarray data may very well follow the same pattern. In the current study, we sought to clarify our understanding of how to optimally analyze data generated using the Invitrogen ProtoArray. Because of the preponderance of relatively low-concentration proteins present on these protein microarrays, we reasoned that interactions of interest in a given experiment were likely to be within the lower-concentration proteins on the array. Toward this end, we herein have devised a novel concentration-dependent analysis method, compared its results to PPI and IRP, two algorithms implemented in the freely available and existing software tool Prospector, and then validated the targets selected by the three different analysis methods using immunological assays.

Despite limitations in the amount of available biologic sample to perform our validation studies, we had sufficient reagents to test 48 of 162 (30%) interactions identified by the different analyses in the CLL data sets, and observed that 30 of 48 (63%) tested protein interactions identified by microarray screening were replicated by immunoprecipitation with *de novo* synthesized protein. The validation rates for PPI, IRP, CDA, and CDA gray zone were 65%, 61%, 72%, and 56%, respectively. This compares favorably with the largest study using human ProtoArray data reported to date, in which only 30% of 100 hits were successfully validated using Western blot analysis.³ Strikingly, although high similarity was observed in validation rates for the individual methods, we observed the far superior validation rate of 92% for proteins that we identified in common between CDA and either PPI or IRP. These data strongly suggest that CDA can be used to focus a list of targets to a highly selected and verifiable group, especially when used in conjunction with complementary analysis methods.

That concentration is an important factor in protein microarray analysis has been previously recognized.^{10,20,24} To test this concept and quantify the results, we performed experimental studies to confirm the effects of concentration correction on hit selection and compared its effects to existing analysis tools. Of note, we observed that correcting for concentration yields a substantial set of additional validated hits. Conceptually, the CDA analysis parallels the DNA microarray analysis method proposed by Kepler,³⁵ where a statistically determined experiment-specific "transcriptional 'core'"; of DNA sequences defines the basis for normalizing signal. CDA is also similar to the concentration correction proposed by Jin et al.²⁴ for enabling duplicate screens by yeast ProtoArrays. We imposed several major criteria for the implementation of CDA, which can be readily adjusted for the needs of a particular

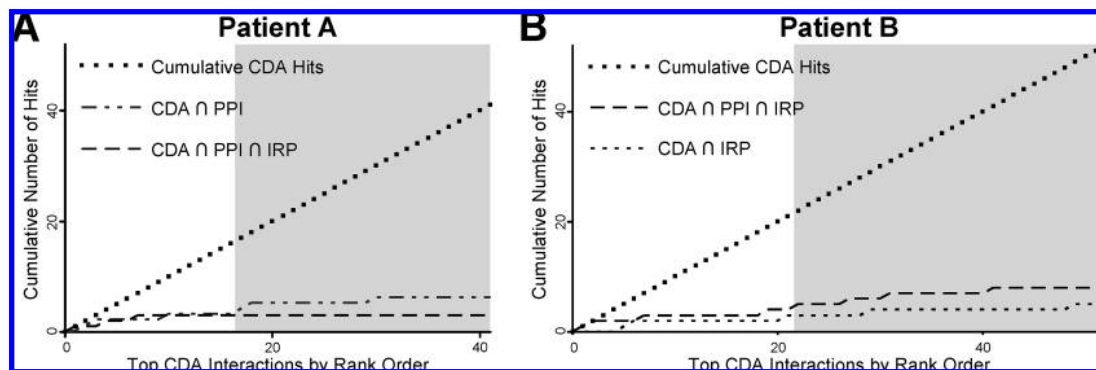


Figure 7. Intersection of hits identified by the three methods, by rank order, using data sets generated from (A) CLL patient A, for whom 19 PPI, 3 IRP, 16 CDA, and 26 CDA gray zone hits were identified, and (B) CLL patient B, for whom 16 PPI, 31 IRP, 21 CDA, and 31 CDA gray zone hits were identified. The CDA hits were sorted by the minimum of the Z_{delta} and Z_{mult} of each of the spots in a pair, which equals the maximal cutoff n which would identify that hit as significant. Because the IRP and PPI methods do not provide a way to rank their identified hits, these are all combined with the top CDA hit at the beginning of the plot and then are combined with subsequent CDA hits as they are added to the plot by rank order. For patient A, IRP hits are a subset of PPI hits, while for patient B, PPI hits are a subset of IRP hits. The gray region depicts where the rank order enters the CDA gray zone.

study. First, the spot neighborhood in CDA was limited to 400 because it provided sufficient data for a good estimate of the local signal distribution, while minimizing the distribution bias. We further restricted the spot neighborhood to 0.75–1.25 of the concentration of the local current spot, balancing the need to limit the local neighborhood to similar spots with the need to provide a smooth estimation of the local mean and standard deviation in the upper and lower 5–10% of spots by concentration. Second, spots that have spot signals significantly different from the local distribution were iteratively removed from local distribution calculations, as previously suggested for DNA microarray data analysis.^{36,37} This is necessary because strong signals can be 50- to 100-fold the level of the local mean, biasing the local distribution and decreasing the significance of other spots. Finally, the spot values themselves, not averages of spot values by concentration, were used.

Our analysis of the 3 data processing methods reveals that CDA and the two Prospector algorithms are highly complementary techniques that sample different areas of the true interactions space. This conclusion is supported by the fact that substantial subsets of the hits identified by the different methods were nonoverlapping. Because the methodology employed by Prospector is not completely described, including details such as how data culling based on quality measures is implemented, a more direct comparison of the methodology to CDA cannot be performed. Nonetheless, it is quite clear that while IRP and PPI are highly related methods, CDA is complementary. We observed that Prospector uses a simple binary threshold for determining significance which yields nonoverlapping data ranges when increased or decreased. When comparing two or more microarrays, as cutoff stringency is varied, nonoverlapping regions of the transformed dataspace are identified as hits. In contrast, the CDA uses data that is transformed nonlinearly before significance testing, and the determination of signal change allows the identification of significant changes in reactivity between samples even if the pretreatment signal was already elevated, as might be meaningful in a biological system. Moreover, it allows the titration of the cutoff level, effectively providing a rank ordering with assigned significance for each protein interaction detected on the microarray.

These differences translate into different advantages and limitations when employing each method as the sole method

of analysis. For example, the advantage of the Prospector cutoff function is that it is easily generalized to any number of simultaneous comparisons. The CDA method, on the other hand, would need adjustment for comparisons among more than two arrays. In addition, because Prospector does not address differences in concentration between printing lots, by design it does not allow combination of experimental data from multiple lots in its analysis. This complicates the comparison of results across different microarray printing runs. Because a main difference between microarrays from different lots is the difference in concentration of protein spots, CDA may potentially allow for the replication of results across microarray printing lots and therefore experiments. This concept will be tested experimentally in our future studies.

The CDA method presented here could be modified in several ways to possibly improve concentration correction and significant protein-binding identification. These include log-transformation of raw data, a common transformation for DNA microarray data;³⁷ use of different thresholds for the two significance-determining functions; combination of the signals or statistics of replicate spots; and selection of the largest significance changes within separate bins by concentration, instead of array-wide selection. Commonly used tests for variation of signal among spot pairs, other tests for signal goodness, and other corrections presented elsewhere were not considered in this study. Finally, the approach to estimating the local signal distribution may be improved. Because the local mean and standard deviation are calculated individually, these curves are not smooth. This particularly affects spots at the extreme ends of concentration, which lack neighbors for a reliable estimation of the local distribution. Moreover, proteins that are spotted at much higher concentrations than the majority of proteins on the array, such as the known autoantigens on the ProtoArray, cannot have a local distribution calculated. To address these issues, one could interpolate the local mean and standard deviations using a smooth curve, or remove all spots with few neighbors from the calculation. Such approaches could also address the slight biases against very low and toward very high concentration spots that are still present in CDA, because spots at high concentration have more neighbors at lower concentrations than at higher concentrations, and *vice versa*, leading to a bias toward increased significance for high- and decreased significance for low-

concentration spots. Although these biases affect all arrays analyzed, they magnify for high- and decrease for low-concentration spots the calculated significance changes.

In closing, we have shown that concentration correction increases the sensitivity of identification of protein microarray results, nearly doubling the number of experimentally validated targets. Our experimental studies demonstrate that a reasonable strategy for selecting a small number of reliable targets would be to look at the intersection of PPI and IRP hits with CDA hits, while a desirable strategy for selecting a broad range of verifiable targets would be to look at the union of PPI, IRP, and CDA hits.

Acknowledgment. We are appreciative of the insightful discussions with Dr. John Quackenbush regarding microarray analyses methods and with Dr. Dawn Mattoon regarding the use of the Invitrogen ProtoArray. We acknowledge the generosity of Drs. David Hill and Mark Vidal for providing reagents for this study; and Jerome Ritz, Stefanie Sarantopoulos, and Ellis Reinherz for critical reading of the manuscript. We thank the clinical transplant team at DFCI for excellent care of patients, and acknowledge the support of the DFCI Pasquarello Tissue Bank for providing clinical samples. O.M. was supported by a Howard Hughes Medical Institute Medical Research Training Fellowship; C.J.W. acknowledges support from the Miles and Eleanor Shore Award, and the NCI (5R21CA115043-2), and is an Early Career Physician-Scientist Awardee of the Howard Hughes Medical Institute.

Supporting Information Available: Supplementary Table 1 lists the significant interactions identified by all three methods, including validation, spot concentration, spot signal pre- and post-treatment, and maximal CDA *n* score which would identify that hit as significant; Supplementary Table 2 lists the ranges of these values found for the spots that either failed or succeeded validation testing, broken down by method identifying the hit and validation outcome. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- MacBeath, G.; Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **2000**, *289* (5485), 1760–3.
- Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. Global analysis of protein activities using proteome chips. *Science* **2001**, *293* (5537), 2101–5.
- Mattoon, D.; Love, B.; Kluger, Y.; Michaud, G.; Schweitzer, B.; Predki, P.; Ritter, G.; Halaban, R. OR65. Melanoma biomarker discovery through serum antibody profiling on protein microarrays. *Clin. Immunol.* **2006**, *119* (Suppl. 1), S28.
- Laliv, P. H.; Menge, T.; Barman, I.; Cree, B. A.; Genain, C. P. Identification of new serum autoantibodies in neuromyelitis optica using protein microarrays. *Neurology* **2006**, *67* (1), 176–7.
- Boutell, J. M.; Hart, D. J.; Godber, B. L.; Kozlowski, R. Z.; Blackburn, J. M. Functional protein microarrays for parallel characterisation of p53 mutants. *Proteomics* **2004**, *4* (7), 1950–8.
- Sundaresh, S.; Doolan, D. L.; Hirst, S.; Mu, Y.; Unal, B.; Davies, D. H.; Felgner, P. L.; Baldi, P. Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics* **2006**, *22* (14), 1760–6.
- Lubomirski, M.; D'Andrea, M. R.; Belkowski, S. M.; Cabrera, J.; Dixon, J. M.; Amaratunga, D. A consolidated approach to analyzing data from high-throughput protein microarrays with an application to immune response profiling in humans. *J. Comput. Biol.* **2007**, *14* (3), 350–9.
- Robinson, W. H.; DiGennaro, C.; Hueber, W.; Haab, B. B.; Kamachi, M.; Dean, E. J.; Fournel, S.; Fong, D.; Genovese, M. C.; de Vegvar, H. E.; Skirner, K.; Hirschberg, D. L.; Morris, R. I.; Muller, S.; Pruijn, G. J.; van Venrooij, W. J.; Smolen, J. S.; Brown, P. O.; Steinman, L.; Utz, P. J. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat. Med.* **2002**, *8* (3), 295–301.
- Michaud, G. A.; Salcius, M.; Zhou, F.; Bangham, R.; Bonin, J.; Guo, H.; Snyder, M.; Predki, P. F.; Schweitzer, B. I. Analyzing antibody specificity with whole proteome microarrays. *Nat. Biotechnol.* **2003**, *21* (12), 1509–12.
- Predki, P. F.; Mattoon, D.; Bangham, R.; Schweitzer, B.; Michaud, G. Protein microarrays: a new tool for profiling antibody cross-reactivity. *Hum. Antibodies* **2005**, *14* (1–2), 7–15.
- He, Q. Y.; Chiu, J. F. Proteomics in biomarker discovery and drug development. *J. Cell Biochem.* **2003**, *89* (5), 868–86.
- Kumble, K. D. Protein microarrays: new tools for pharmaceutical development. *Anal. Bioanal. Chem.* **2003**, *377* (5), 812–9.
- Gunawardana, C. G.; Diamandis, E. P. High throughput proteomic strategies for identifying tumour-associated antigens. *Cancer Lett.* **2007**, *249* (1), 110–9.
- Luscombe, N. M.; Royce, T. E.; Bertone, P.; Echols, N.; Horak, C. E.; Chang, J. T.; Snyder, M.; Gerstein, M. ExpressYourself: A modular platform for processing and visualizing microarray data. *Nucleic Acids Res.* **2003**, *31* (13), 3477–82.
- Colantuoni, C.; Henry, G.; Zeger, S.; Pevsner, J. SNOMAD (Standardization and Normalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics* **2002**, *18* (11), 1540–1.
- Saeed, A. I.; Bhagabati, N. K.; Braisted, J. C.; Liang, W.; Sharov, V.; Howe, E. A.; Li, J.; Thiagarajan, M.; White, J. A.; Quackenbush, J. TM4 microarray software suite. *Methods Enzymol.* **2006**, *411*, 134–93.
- Brusic, V.; Marina, O.; Wu, C. J.; Reinherz, E. L. Proteome informatics for cancer research: from molecules to clinic. *Proteomics* **2007**, *7* (6), 976–91.
- Royce, T. E.; Rozowsky, J. S.; Luscombe, N. M.; Emanuelsson, O.; Yu, H.; Zhu, X.; Snyder, M.; Gerstein, M. B. Extrapolating traditional DNA microarray statistics to tiling and protein microarray technologies. *Methods Enzymol.* **2006**, *411*, 282–311.
- Robinson, W. H.; Fontoura, P.; Lee, B. J.; de Vegvar, H. E.; Tom, J.; Pedotti, R.; DiGennaro, C. D.; Mitchell, D. J.; Fong, D.; Ho, P. P.; Ruiz, P. J.; Mavarakis, E.; Stevens, D. B.; Bernard, C. C.; Martin, R.; Kuchroo, V. K.; van Noort, J. M.; Genain, C. P.; Amor, S.; Olsson, T.; Utz, P. J.; Garren, H.; Steinman, L. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nat. Biotechnol.* **2003**, *21* (9), 1033–9.
- Zhu, X.; Gerstein, M.; Snyder, M. ProCAT: a data analysis approach for protein microarrays. *Genome Biol.* **2006**, *7* (11), R110.
- Horn, S.; Lueking, A.; Murphy, D.; Staudt, A.; Gutjahr, C.; Schulte, K.; Konig, A.; Landsberger, M.; Lehrach, H.; Felix, S. B.; Cahill, D. J. Profiling humoral autoimmune repertoire of dilated cardiomyopathy (DCM) patients and development of a disease-associated protein chip. *Proteomics* **2006**, *6* (2), 605–13.
- Sreekumar, A.; Nyati, M. K.; Varambally, S.; Barrette, T. R.; Ghosh, D.; Lawrence, T. S.; Chinnaiyan, A. M. Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res.* **2001**, *61* (20), 7585–93.
- Invitrogen Home page. <http://www.invitrogen.com/protoarray>.
- Jin, F.; Hazbun, T.; Michaud, G. A.; Salcius, M.; Predki, P. F.; Fields, S.; Huang, J. A pooling-deconvolution strategy for biological network elucidation. *Nat. Methods* **2006**, *3* (3), 183–9.
- Hu, S.; Li, Y.; Liu, G.; Song, Q.; Wang, L.; Han, Y.; Zhang, Y.; Song, Y.; Yao, X.; Tao, Y.; Zeng, H.; Yang, H.; Wang, J.; Zhu, H.; Chen, Z. N.; Wu, L. A protein chip approach for high-throughput antigen identification and characterization. *Proteomics* **2007**, *7* (13), 2151–61.
- Olle, E. W.; Sreekumar, A.; Warner, R. L.; McClintock, S. D.; Chinnaiyan, A. M.; Bleavins, M. R.; Anderson, T. D.; Johnson, K. J. Development of an internally controlled antibody microarray. *Mol. Cell. Proteomics* **2005**, *4* (11), 1664–72.
- Invitrogen, ProtoArray Prospector v4.0 User Guide, Version E, 2006.
- Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screening* **1999**, *4* (2), 67–73.
- Liang, F.; Matrubutham, U.; Parvizi, B.; Yen, J.; Duan, D.; Mirchandani, J.; Hashima, S.; Nguyen, U.; Ubil, E.; Loewenheim, J.; Yu, X.; Sipes, S.; Williams, W.; Wang, L.; Bennett, R.; Carrino, J. ORFDB: an information resource linking scientific content to a high-quality Open Reading Frame (ORF) collection. *Nucleic Acids Res.* **2004**, *32* (Database issue), D595–9.
- Zuo, D.; Mohr, S. E.; Hu, Y.; Taycher, E.; Rolfs, A.; Kramer, J.; Williamson, J.; LaBaer, J. PlasmID: a centralized repository for plasmid clone information and distribution. *Nucleic Acids Res.* **2007**, *35* (Database issue), D680–4.

- (31) Rual, J. F.; Hirozane-Kishikawa, T.; Hao, T.; Bertin, N.; Li, S.; Dricot, A.; Li, N.; Rosenberg, J.; Lamesch, P.; Vidalain, P. O.; Clingsmith, T. R.; Hartley, J. L.; Esposito, D.; Cheo, D.; Moore, T.; Simmons, B.; Sequerra, R.; Bosak, S.; Doucette-Stamm, L.; Le Peuch, C.; Vandenhaute, J.; Cusick, M. E.; Albala, J. S.; Hill, D. E.; Vidal, M. Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.* **2004**, *14* (10B), 2128–35.
- (32) Lennon, G.; Auffray, C.; Polymeropoulos, M.; Soares, M. B. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **1996**, *33* (1), 151–2.
- (33) Bussey, K. J.; Kane, D.; Sunshine, M.; Narasimhan, S.; Nishizuka, S.; Reinhold, W. C.; Zeeberg, B.; Ajay, W.; Weinstein, J. N. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *GenomeBiology* **2003**, *4* (4), R27.
- (34) Quackenbush, J. Microarray analysis and tumor classification. *N. Engl. J. Med.* **2006**, *354* (23), 2463–72.
- (35) Kepler, T. B.; Crosby, L.; Morgan, K. T. Normalization and analysis of DNA microarray data by self-consistency and local regression. *GenomeBiology* **2002**, *3* (7), research0037.
- (36) Yang, I. V.; Chen, E.; Hasseman, J. P.; Liang, W.; Frank, B. C.; Wang, S.; Sharov, V.; Saeed, A. I.; White, J.; Li, J.; Lee, N. H.; Yeatman, T. J.; Quackenbush, J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *GenomeBiology* **2002**, *3* (11), research0062.
- (37) Quackenbush, J. Microarray data normalization and transformation. *Nat. Genet.* **2002**, *32* (Suppl.), 496–501.

PR700892H

Patient	Immuno- globulin	Validation Outcome	PPI	IRP	CDA hit	CDA gray zone	CDA <i>n</i>	Spot 1 Concent ration	Spot 2 Concent ration	Spot 1 Spot Signal Pre
A		Passed	yes	yes	yes		39,40	2730	2740	1197
A		Passed	yes		yes		28,31	849	940	1624
A		Passed	yes		yes		13,93	760	763	1375
A		Passed			yes		8,67	50,9	51,1	-17
A		Passed	yes	yes	yes		8,43	3270	3270	4234
A					yes		6,58	1150	1250	1231
A		Passed	yes	yes	yes		6,51	3250	3270	5045
A					yes		6,49	1990	2100	1315
A		Passed			yes		5,99	885	942	747
A			yes		yes		5,90	1620	1710	2602
A					yes		5,68	271	273	1712
A					yes		5,51	263	277	1281
A					yes		5,48	182	187	347
A		Failed			yes		5,33	585	618	2340
A		Passed	yes		yes		5,24	109	123	2643
A		Passed			yes		5,02	490	550	900
A		Passed	yes			yes	4,82	241	271	3015
A						yes	4,75	124	130	1026
A						yes	4,74	53,4	57,6	962
A						yes	4,72	274	282	1302
A						yes	4,66	211	219	1896
A						yes	4,43	664	702	754
A						yes	4,18	212	238	1855
A						yes	4,14	74	78,6	1214
A						yes	4,02	50,8	52,5	714
A						yes	4,01	236	239	1869
A						yes	3,96	951	999	1996
A						yes	3,96	400	430	2292
A						yes	3,82	71,1	74,7	575
A						yes	3,79	184	191	1749
A		Passed	yes			yes	3,76	189	198	3073
A						yes	3,74	1670	1720	1259
A						yes	3,57	47,5	49,4	427
A						yes	3,56	484	503	769
A						yes	3,52	299	311	1557
A						yes	3,52	1660	1770	2724
A						yes	3,43	446	466	289
A						yes	3,32	342	389	2742
A						yes	3,20	3230	3260	1507
A						yes	3,17	341	344	1201
A		Failed				yes	3,16	263	290	2340
A						yes	3,10	71,6	73,7	1290
A		Failed	yes				2,55	749	788	4289
A		Passed	yes				2,08	1500	1610	5905
A		Failed	yes				1,94	1390	1430	7809
A			yes				1,75	1670	1690	5077

A	Passed	yes			1,70	2270	2440	7732
A		yes			1,62	985	1060	6158
A		yes			1,56	2760	2800	6109
A		yes			1,42	3270	3270	6806
A		yes			1,37	3270	3270	5927
A		yes			1,35	2920	3130	5558
B	Passed	yes	yes	yes	30,68	1150	1260	578
B	Passed	yes	yes	yes	19,59	2070	2150	1272
B	Failed			yes	14,39	966	1040	442
B	Passed			yes	13,12	881	940	634
B	Passed		yes	yes	11,97	3430	3460	280
B	Failed	yes	yes	yes	11,68	3330	3450	107
B	Passed		yes	yes	10,59	737	790	894
B	Passed			yes	8,35	1350	1500	773
B	Passed			yes	8,33	98,3	102	229
B	Passed			yes	7,68	564	610	794
B	Failed			yes	7,28	1080	1150	260
B	Failed			yes	6,77	2290	2460	481
B				yes	6,66	1310	1470	839
B	Failed			yes	6,65	3270	3270	618
B				yes	6,62	783	848	526
B	Failed			yes	6,44	2420	2680	631
B	Passed			yes	6,43	257	282	649
B	Passed			yes	6,39	760	763	497
B	Passed	yes	yes	yes	5,63	3270	3270	2516
B	Passed		yes	yes	5,50	2420	2440	1514
B	Failed			yes	5,30	549	615	731
B				yes	4,98	1420	1430	993
B				yes	4,98	1720	1840	437
B	Failed	yes	yes	yes	4,97	564	594	2128
B	Failed		yes	yes	4,71	2540	2620	1825
B		yes	yes	yes	4,70	2920	3130	2122
B				yes	4,69	721	750	179
B				yes	4,43	660	660	427
B				yes	4,42	2150	2550	1001
B				yes	4,35	1390	1450	1009
B				yes	4,22	2100	2250	781
B				yes	4,14	664	702	467
B				yes	4,08	111	115	359
B				yes	3,88	1040	1140	608
B	Passed	yes	yes	yes	3,64	3250	3270	2384
B		yes	yes	yes	3,56	3270	3270	2269
B	Failed			yes	3,48	1020	1120	1176
B				yes	3,47	2460	2500	1576
B				yes	3,41	448	481	837
B				yes	3,34	706	823	382
B	Passed			yes	3,30	29,5	31,7	611
B				yes	3,27	1020	1070	948
B				yes	3,26	451	516	781
B				yes	3,25	184	186	224
B				yes	3,21	113	118	496

B		Passed		yes	3,20	117	120	1075
B				yes	3,15	146	159	797
B			yes	yes	3,13	3270	3270	2037
B				yes	3,11	560	570	733
B				yes	3,09	1820	1920	606
B				yes	3,08	409	445	505
B				yes	3,03	2390	2640	1213
B		Failed	yes	yes	2,99	3270	3360	3326
B			yes	yes	2,93	3270	3270	2634
B		Failed	yes	yes	2,38	1780	1880	3264
B		Failed		yes	2,01	673	707	2670
B				yes	1,83	1270	1370	3470
B				yes	1,78	817	884	3241
B		Failed	yes	yes	1,75	3340	3380	4334
B	yes			yes	1,67	425	430	4560
B		Passed		yes	1,66	2560	2730	4472
B	yes		yes	yes	1,59	66,5	70,2	5694
B	yes		yes	yes	1,45	481	494	5207
B		Passed	yes	yes	1,44	1660	1770	4645
B		Passed		yes	1,41	1980	2130	4816
B				yes	1,37	1670	1690	3605
B		Passed		yes	1,32	1460	1540	4074
B				yes	1,18	3050	3080	4028
B		Failed	yes	yes	1,15	769	857	5481
B		Failed		yes	1,08	273	292	5049

Spot 1	Spot 2	Spot 2
Spot	Spot	Spot
Signal	Signal	Signal
Post	Pre	Post
51988	1247	53175
25214	1503	28841
13444	1436	13149
6166	-141	6245
39723	4370	34146
7772	1159	7643
33486	5222	36723
9623	1216	10258
6723	688	6689
11462	2761	12171
6263	1772	6166
5677	1097	5939
4991	176	4802
9591	2190	9883
12880	2169	13434
5758	666	5856
12929	2682	13788
4524	900	5170
3841	947	4832
5198	1560	4763
6072	1903	6086
5454	800	6204
5480	1835	5569
4216	1585	4938
3351	846	5098
5422	1919	6172
5823	1596	5674
6892	2302	6976
3684	672	4282
4940	1723	4539
10032	3169	10377
5963	1336	6555
3081	397	3908
4247	727	4523
4100	1746	4910
8028	2601	8628
3850	273	5008
7671	2739	7495
6808	1626	6712
3587	1151	3602
6140	2248	7838
4095	1607	3929
10218	4017	10491
15421	6046	15353
17564	7975	18927
10711	5018	12234

17315	7560	17227
11862	5960	12613
11807	6233	12572
12603	6645	12608
11798	5979	10629
10467	5938	10657
6221	506	6338
12388	1252	12040
2805	491	2929
3585	630	3599
5124	373	6095
11009	128	12480
7017	776	4670
3284	810	3145
1125	159	1147
3273	729	3362
1711	211	1680
2718	455	2488
3277	953	3454
2976	634	3062
1713	497	2377
2589	765	2648
2306	614	2238
1739	538	1822
11178	2606	11612
4722	1492	4673
2287	759	2232
2798	949	2822
1777	450	1896
7751	2209	8150
5338	1717	5122
7669	2112	7296
1282	170	1368
1293	373	1240
2281	997	2187
2775	1005	2614
1929	760	1815
1510	445	1252
765	325	762
1257	570	1233
8560	2513	7397
6365	2194	6805
2777	1311	3296
3579	1613	3436
1808	892	1934
1032	437	1173
1212	525	1164
1976	903	1860
1542	755	1538
656	166	753
1008	384	963

2557	998	2333
2222	909	2264
4945	1918	4978
1347	669	1374
1371	618	1361
1043	609	1010
2100	1239	2217
10129	3576	9706
6703	2690	6599
6545	3242	6638
4808	2780	4850
5271	3558	5326
4920	2373	4545
7491	4241	7233
5391	4476	5719
5920	3921	5632
7120	5660	7347
6162	4990	6373
6756	4579	7417
5484	4781	5553
5127	3961	5322
4935	4181	5176
4471	4050	4467
5833	5076	6167
4500	4903	4876

Patient	Validation Outcome	PPI or IRP hit	CDA hit	Group Size	CDA score: median (low-high)	Concentration: median (low-high)
A	Passed	yes	yes	6	11.2 (5.2-39.4)	1835 (109-3270)
A	Failed	yes	yes	0	n/a	n/a
A	Passed	yes		4	2.9 (1.7-4.8)	886 (189-2440)
A	Failed	yes		2	2.2 (1.9-2.6)	1089 (749-1430)
A	Passed		yes	3	6.0(5.0-8.7)	530 (51-942)
A	Failed		yes	1	5,3	602 (585,618)
B	Passed	yes	yes	6	11.3 (5.5-30.7)	2285 (737-3460)
B	Failed	yes	yes	1	11,7	3390 (3330,3450)
B	Passed	yes		5	1.4 (1.3-3.6)	2055 (1460-3270)
B	Failed	yes		8	2.2 (1.1-5.0)	1319 (273-3380)
B	Passed		yes	6	8.0 (6.4-13.1)	642 (98-1500)
B	Failed		yes	6	6.7 (5.3-14.4)	1720 (549-3270)

**Post Spot Signal:
median (low-high)**

31164 (12880-53175)
n/a
14576 (10032-17315)
14028 (10218-18927)
6205 (5758-6723)
9737 (9591,9883)
6280 (4670-12388)
11745 (11009,12480)
5776 (4935-8560)
6356 (4500-10129)
2726 (1125-3599)
2619 (1680-3062)