Automated semiquantitative analysis of protein macroarrays

*Chin Hong Ooi, Nam-Trung Nguyen, Gregor S Kijanka**

Queensland Micro-Nanotechnology Centre, Griffith University, Nathan, QLD 4111, Australia

*Author to whom correspondence should be addressed. Email: g.kijanka@griffith.edu.au

KEYWORDS Protein arrays, macroarrays, proteomics, antibody, semiquantitative analysis

ABSTRACT Protein arrays are systematically arranged, large collections of annotated proteins on planar surfaces commonly used for the characterisation of protein binding events against a wide range of possible probes. These may include analyses of protein-protein, peptide-protein, enzyme-substrate or antibody-antigen interactions from simple reagents to complex mixtures. Absence of appropriate image analysis and data processing software may bestow a substantial hurdle limiting the uptake of protein arrays in research. We developed a first, automated semiquantitative open source software package for the analysis of widely used protein macroarrays. The software allows accurate single array and inter-array comparative studies through the tackling of intra-array inconsistencies arising from experimental disparities. The innovative and automated image analysis process includes adaptive positioning, background identification and subtraction, removal of null signals, robust statistical analysis, and protein pair validation. The normalized values allow a convenient semiquantitative data analysis of different

samples or timepoints, enabling accurate characterisation of sample series to identify relative changes for instance in clinical samples in response to diseases and treatment.

INTRODUCTION

Protein arrays are a well-established proteomic tool for the simultaneous analysis of thousands of interaction partners, generally manufactured in an microarray format through immobilization of purified proteins on chemically modified microscope slides [1]. An alternative to the microarray format are protein macroarrays produced by printing of annotated libraries of *E.coli* clones, expressing recombinant human protein, on large 22×22cm polyvinylidene fluoride (PVDF) membranes [2, 3]. Since their introduction in the early 2000' the protein macroarrays have been frequently used with around 100 published studies in a wide range of applications from protein-protein [4, 5], peptide-protein [6, 7], enzyme-substrate [8, 9] and posttranslational modification interaction studies [10, 11], to antibody specificity validation [12, 13], antibody target discovery [14, 15], antibody isotyping [16-18] and clinical autoantibody screening [11, 16, 19]. The extensive usage of the protein microarrays can be attributed to two innate characteristics unique to the platform: i) *E.coli* expression clones are spotted directly on PVDF membranes, ensuring consistent protein concentrations intrinsic to each individual expression clone, thereby avoiding the need for cumbersome large-scale protein purification and characterisation procedures essential for the generation of most protein microarray formats; and ii) each individual colony spot on the macroarray comprises of a single recombinant human protein and a collection of all *E.coli* proteins, thereby providing a natural blocking background consistent across the entire array and hence ensuring excellent experimental signal to noise ratios [20].

While users of protein microarrays may draw on dedicated commercial and open source microarray analysis software packages [21, 22], such packages are not suitable for protein macroarrays due to the large array size and associated image resolution characteristics [23, 24]. Protein macroarray analysis options are thus far limited to image data processing of subsections of the array [25, 26] or universal manufacturer-provided scanner software packages [10]. More dedicated software packages such as VisualGrid (GPC Biotech) [19, 27] or Aida Image Analyser (Raytest) [5, 28] rely either on operator-based decision making or fixed spot diameter measurements which are not suitable for semiquantitative comparative studies. Here we aim to develop an automated open source software package dedicated to semiquantitative analysis of protein macroarrays. We use MathWorks MATLAB version 2019b as the platform. The code is available for use in the supplementary information section.

EXPERIMENTAL PROCEDURES

Protein Macroarrays: HexSelect protein macroarrays (Engine GmbH, Germany) were prepared as previously described [19]. Briefly, protein macroarrays were incubated with diluted volunteer serum (1:100) for 16 h. Mouse anti-human IgG antibody (GG-7, Sigma-Aldrich), alkaline phosphatase (AP)-conjugated goat anti-mouse IgG antibody (A1418, Sigma-Aldrich) and AttoPhos substrate (S1000; Promega) were used as detection reagents. The arrays were scanned using GE Typhoon FLA 9000 Gel Imaging Scanner (GE Healthcare, Chicago, IL, USA).

MATLAB Programming: MATLAB code is written specifically for the HexSelect protein macroarrays imaged using a 16-bit scanner at a resolution of 10 pixels/mm. The user interface built into the code requires MATLAB version 2019b or later to work properly. We used uncompressed TIFF files converted from the scanner RAW file without any compression or

noise removal for the best results. All images were checked to ensure that there was no pixel saturation, i.e. all pixel values were well within the linear dynamic range of 0 to 65,535. These macroarrays contain 57,600 “dots” that are grouped into 2,304 “cells”. Each cell consists of a 5 X 5 array of 25 dots, including the marker dot at the centre and 12 pairs of clones in a specific configuration as previously shown [19]. Each dot is represented by 9 X 9 pixels on the image.

RESULTS AND DISCUSSION

Image loading and aligning of the array

A protein macroarray TIFF images was loaded from the MATLAB folder via a user dialogue. The image may be mirrored depending on the side of array that was scanned. The images used in this study were all the same size at 2,250 X 2,250 pixels.

An accurate dot positioning is critical for large protein arrays as minute distortions present in the substrate accumulate throughout the entire array. Nevertheless, the large dimensions facilitate accurate physical placement in terms of orientation and thus substantially reduce positioning errors arising from rotation. Array scanners tend to sacrifice spatial resolution to improve intensity resolution and signal-to-noise ratio (SNR) by using charge-coupled device (CCD) cameras with relatively large pixels. Therefore, positioning accuracy is paramount to reliable data.

Physical distortion of the array substrate is non-linear and heterogenous. As such, the ubiquitous 3-point interpolation positioning method that assumes perfect grid-like array structure fails to accurately determine dot locations in large arrays. To address this issue, we implement a two-stage positioning method that includes a 4-point interpolation positioning method followed by a

secondary adaptive position refinement aided by marker dots located at the centre of each cell. The coordinates of the marker dots at the four corners of the array are the only user inputs required by our method.

The 4-point positioning method forms a quadrilateral with evenly interpolated spacing as an initial approximation to locate marker dots. As the dots were printed black with high contrast, the centre of each marker dot is accurately determined via simple pixel intensity thresholding. Once determined, the pixel coordinates of the centre of each marker dot form the centre of each cell. **Figure 1** shows positions of the marker dots acquired using this method.

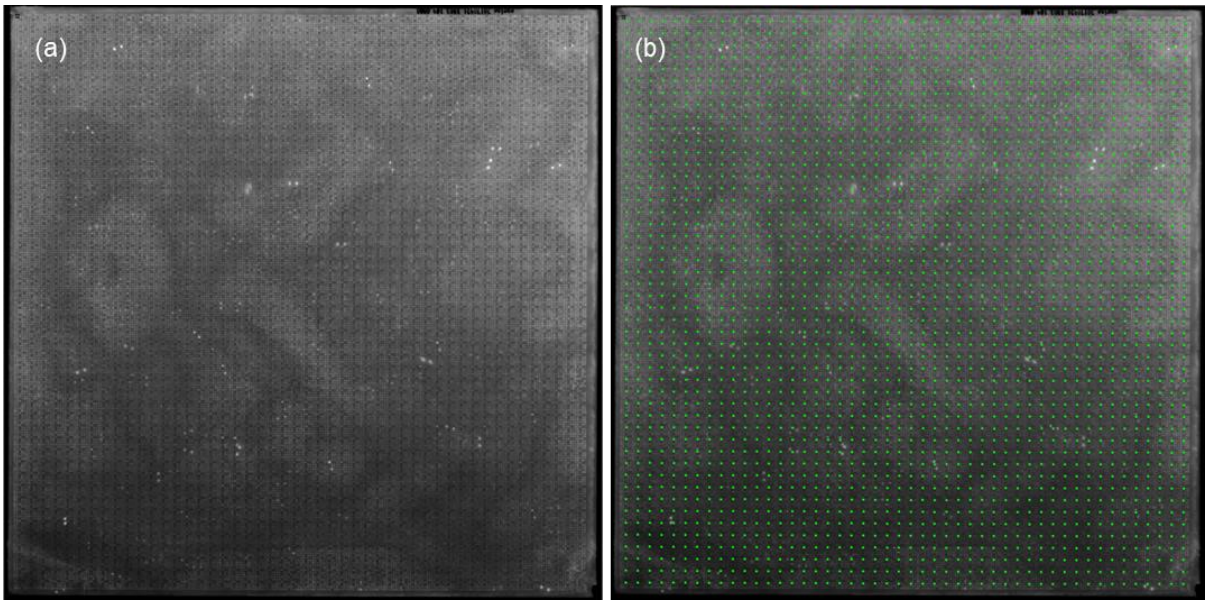

Figure 1. Determining the positions of the marker dots. (a) Unprocessed image of the protein array. The heterogenous background is clearly visible. (b) Image overlaid with green marker positions matching the positions of printed black marker dots.

Reconstruction of array

The array is next reconstructed such that the image analysis is only applied to regions of interest (2,160 X 2,160 pixels). Each cell is redefined by using the marker dot centre as datum, such that the cell area is ±22 pixels in the horizontal and vertical direction from the centre pixel. To check the accuracy of the marker dot position and subsequent reconstruction, the code samples 6 X 6 cells that are evenly spaced across the entire array. A sample result is shown in **Figure 2**.

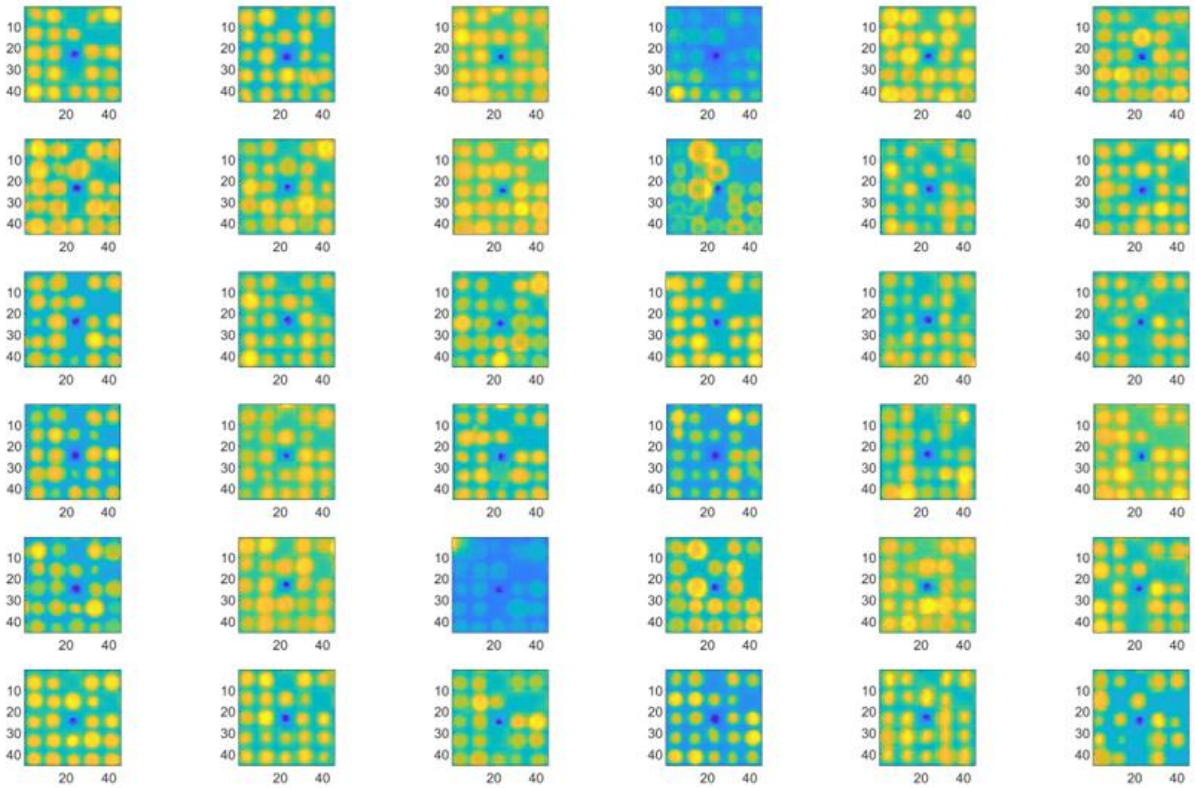

Figure 2. Sample cells to illustrate positioning accuracy throughout the entire array. The colour is applied by MATLAB and normalised according to pixel intensity. The marker dots are visible at the centre of each cell in dark blue as they have the lowest pixel intensities.

Removal of null results

The protein arrays contain null results that include marker dots and dots that do not produce any signal, hereby referred to as blank dots. Each cell can have up to 2 pairs of blank dots that should be excluded from subsequent statistical analyses. Marker dots were removed based on previously known positions whereas blank dots were removed based on divergence of the pixel intensity field. Dots with signals form sinks in the divergence field. **Figure 3** shows the position of the blank dots and marker dots throughout the entire array.

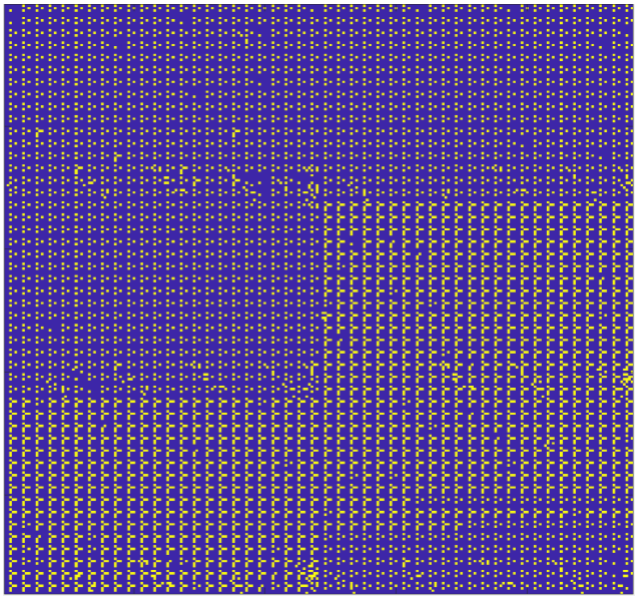

Figure 3. Detected blank dots in the array. Visible straight boundaries between regions of different blank dot densities suggest artefacts during array production.

Background subtraction

We do not include any user input for background subtraction to improve reproducibility. To quantify the signal strength of each dot, the background pixel offset and gain should be subtracted. For large arrays with human-derived samples, local variations in offset and gain are

significant. Variations were observed even within one cell. As such, we assume that offsets are only homogenous at the dot level. Within each dot, the background offset pixel value is defined as the intensity of the pixel(s) that expresses no fluorescence. As such, this corresponds to the pixel with minimum intensity. Dots with extremely strong signal can fill the entire dot space with fluorescent signal. In these cases, background pixel values would be abnormally high and detected as outliers. Outlier background values will be replaced by local median background value instead.

The locality of a dot is defined as the sample of dots contained within a radius of 4 dots from the dot of interest, similar to a moving window. Selection of radius size takes into account of competing factors. A small radius will have a higher homogeneity of pixel gain but suffers from statistically unreliable small samples. A 3-dot radius completely covers the area of one cell with 29 dots. At the worst case, only 4 dots lie outside of the cell. As such, a 3-dot radius locality can be susceptible to errors within a cell. A 4-dot radius yields 49 dots which includes 24 dots outside of the cell at worst. As such, the 4-dot radius is used in our analysis. **Figure 4** shows the comparison between a 3-dot and 4-dot radius locality.

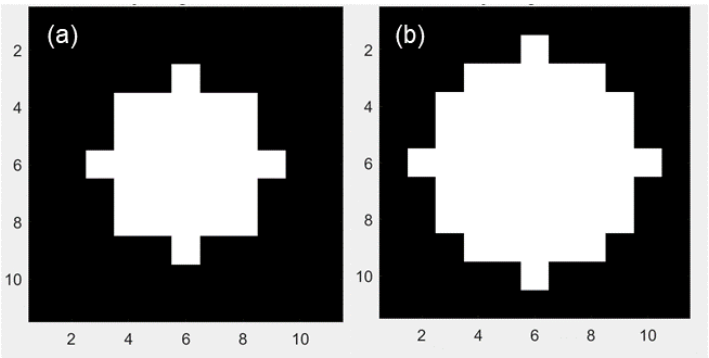

Figure 4. (a) 3-dot and (b) 4-dot radius localities in an array.

Next, the entire array is reworked by subtracting each pixel with its corresponding offset, **Figure 5**. Once the offset is removed, only the variable gain component remains. The gain component can be determined if a known reference value is available. Since the arrays do not contain a reference value, only relative gain values can be determined.

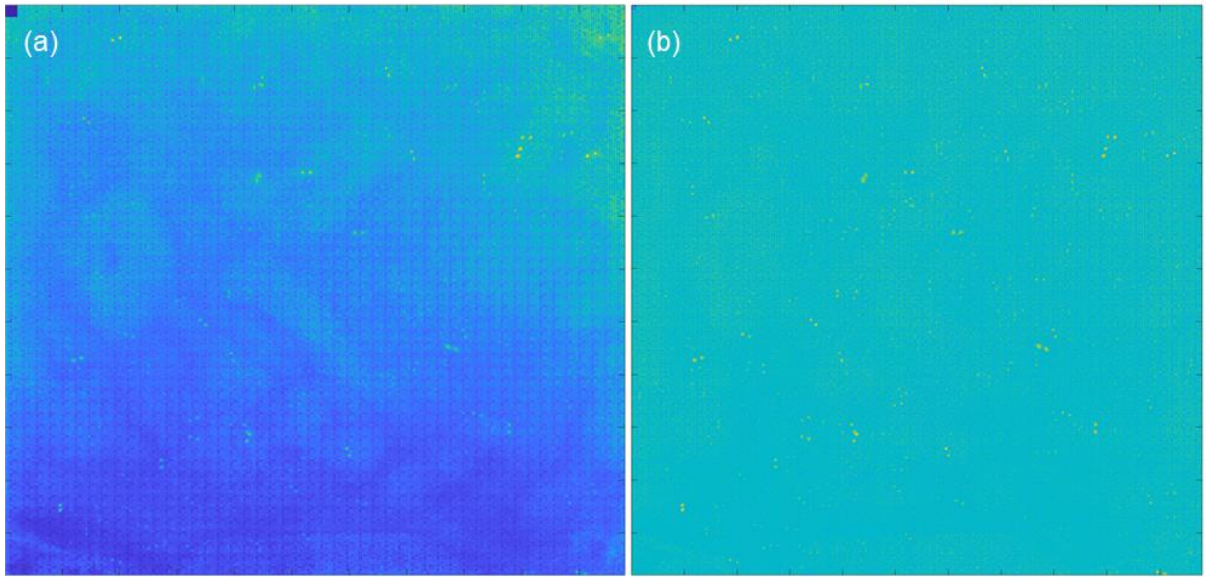

Figure 5. (a) Before and (b) after background subtraction. Note that our background subtraction method did not suppress the strong signals within the array.

Pixel value summation

Pixel values within each dot is summed to obtain the total pixel intensity. Unlike other platforms, we do not require a shape mask to determine region of interest within the dot. In our case, the entire dot (9 X 9 pixels) will be included in the dot pixel sum since the background pixel value is negligible after the background subtraction process. The result is a 240 X 240 numerical array.

Median absolute deviation (MAD) calculation

Since there is no reference to an absolute value, we define overexpression as a dot that has significantly higher intensity than dots in its locality. Since an overexpression is a statistical outlier, we measure the degree of spread of each dot from the local centre of distribution in terms of pixel intensity. The median absolute deviation (MAD) is calculated for each locality. Each dot of interest will yield an absolute deviation from median which is normalised with respect to its local MAD. As such, the local gain component is eliminated. The result is the number of MADs from median which is reported as a relative quantification of expression levels. Since the local offset and gain components were removed, the number of MADs from median can be used to compare levels of expressions across arrays from different samples and time points.

The number of MADs from median for all 57,600 dots are stored in a linearised array with their corresponding x and y coordinate to facilitate matching with the protein library. **Figure 6(a)** illustrates the outliers detected based on the customisable MAD threshold > 3. As the output is a generic numerical array, the results can be analysed or filtered in Microsoft Excel or MATLAB. These results can be stacked for comparison or longitudinal study purposes.

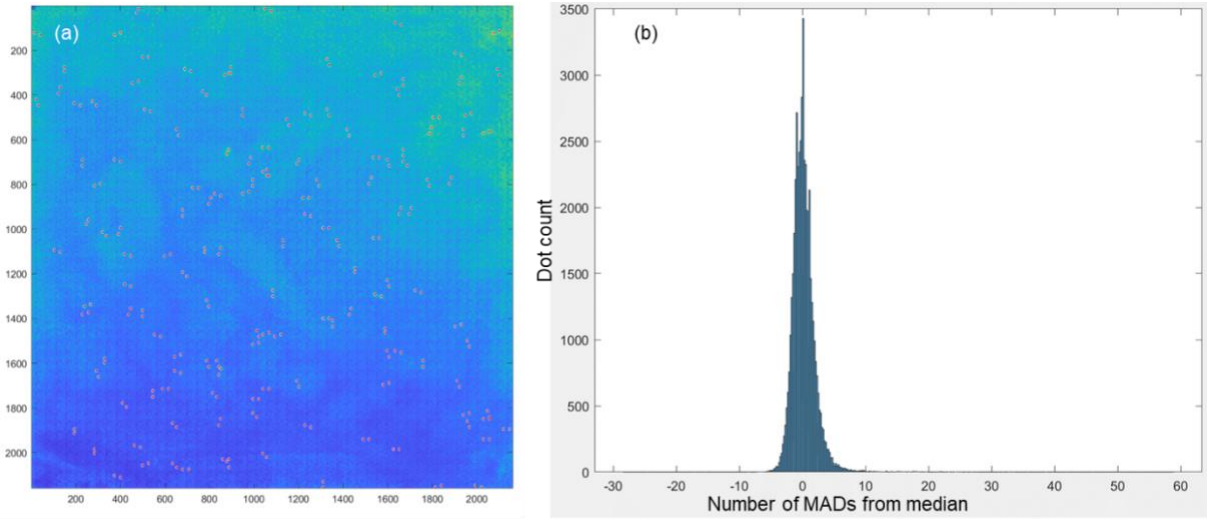

Figure 6. Output results. (a) Dot pairs with MAD > 3 highlighted in red. (b) Histogram of dot count vs. number of MADs from median. Note that this array contains extremely strong signals that have MAD > 50.

CONCLUSIONS

We developed a first, automated semiquantitative open source software package for the analysis of widely used protein macroarrays. The software allows accurate single array and inter-array comparative studies through the tackling of intra-array inconsistencies arising from experimental disparities. The innovative and automated image analysis process includes adaptive positioning, background identification and subtraction, removal of null signals, robust statistical analysis, and protein pair validation. The normalized values allow a convenient semiquantitative data analysis of different samples or timepoints, enabling accurate characterisation of sample series to identify relative changes for instance in clinical samples in response to diseases and treatment. The associated code is available for use in the supplementary information section.

ASSOCIATED CONTENT

Supplementary code file

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was performed in part at the Queensland node of the Australian National Fabrication Facility, a company established under the National Collaborative Research Infra- structure

Strategy to provide nano- and microfabrication facilities for Australia's researchers. N.T.N and C.H.O acknowledge funding support from The Australian Research Council, project DP220100261. GSK acknowledges funding support from Retention of High Performing Research Fellows Grant SDV2723 Griffith University.

REFERENCES

1. Romanov, V., et al., *A critical comparison of protein microarray fabrication technologies.* Analyst, 2014. **139**(6): p. 1303-26.
2. Bussow, K., et al., *A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library.* Nucleic Acids Res, 1998. **26**(21): p. 5007-8.
3. Bussow, K., et al., *A human cDNA library for high-throughput protein expression screening.* Genomics, 2000. **65**(1): p. 1-8.
4. Ludes-Meyers, J.H., et al., *WWOX binds the specific proline-rich ligand PPXY: identification of candidate interacting proteins.* Oncogene, 2004. **23**(29): p. 5049-55.
5. Grelle, G., et al., *Identification of VCP/p97, carboxyl terminus of Hsp70-interacting protein (CHIP), and amphiphysin II interaction partners using membrane-based human proteome arrays.* Mol Cell Proteomics, 2006. **5**(2): p. 234-44.
6. Larkin, D., et al., *ICln, a novel integrin alphaIIbbeta3-associated protein, functionally regulates platelet activation.* J Biol Chem, 2004. **279**(26): p. 27286-93.
7. Arthur, J.F., et al., *TNF receptor-associated factor 4 (TRAF4) is a novel binding partner of glycoprotein Ib and glycoprotein VI in human platelets.* J Thromb Haemost, 2011. **9**(1): p. 163-72.
8. Lee, J. and M.T. Bedford, *PABP1 identified as an arginine methyltransferase substrate using high-density protein arrays.* EMBO Rep, 2002. **3**(3): p. 268-73.
9. de Graaf, K., et al., *Characterization of cyclin L2, a novel cyclin with an arginine/serine-rich domain: phosphorylation by DYRK1A and colocalization with splicing factors.* J Biol Chem, 2004. **279**(6): p. 4612-24.
10. Pless, O., et al., *A differential proteome screening system for post-translational modification-dependent transcription factor interactions.* Nat Protoc, 2011. **6**(3): p. 359-64.
11. Gronwall, C., et al., *A Comprehensive Evaluation of the Relationship Between Different IgG and IgA Anti-Modified Protein Autoantibodies in Rheumatoid Arthritis.* Front Immunol, 2021. **12**: p. 627986.
12. Kijanka, G., et al., *Rapid characterization of binding specificity and cross-reactivity of antibodies using recombinant human protein arrays.* J Immunol Methods, 2009. **340**(2): p. 132-7.
13. Lemass, D., R. O'Kennedy, and G.S. Kijanka, *Referencing cross-reactivity of detection antibodies for protein array experiments.* F1000Res, 2016. **5**: p. 73.
14. Kijanka, G., et al., *Defining the molecular target of an antibody derived from nuclear extract of Jurkat cells using protein arrays.* Anal Biochem, 2009. **395**(2): p. 119-24.

15. Kleine, A.D. and B. Reuss, *Interactions of Antibodies to the Gram-Negative Gastric Bacterium Helicobacter pylori with the Synaptic Calcium Sensor Synaptotagmin 5, Correlate to Impaired Vesicle Recycling in SiMa Human Neuroblastoma Cells.* J Mol Neurosci, 2021. **71**(3): p. 481-505.
16. Horn, S., et al., *Profiling humoral autoimmune repertoire of dilated cardiomyopathy (DCM) patients and development of a disease-associated protein chip.* Proteomics, 2006. **6**(2): p. 605-13.
17. Islam Roney, M.M.S., et al., *Isotypic analysis of anti-p53 serum autoantibodies and p53 protein tissue phenotypes in colorectal cancer.* Hum Pathol, 2022. **128**: p. 1-10.
18. Roney, M.S.I., et al., *IgM and IgA augmented autoantibody signatures improve early-stage detection of colorectal cancer prior to nodal and distant spread.* Clin Transl Immunology, 2021. **10**(9): p. e1330.
19. Kijanka, G., et al., *Human IgG antibody profiles differentiate between symptomatic patients with and without colorectal cancer.* Gut, 2010. **59**(1): p. 69-78.
20. Kijanka, G. and D. Murphy, *Protein arrays as tools for serum autoantibody marker discovery in cancer.* J Proteomics, 2009. **72**(6): p. 936-44.
21. Da Gama Duarte, J., et al., *PMA: Protein Microarray Analyser, a user-friendly tool for data processing and normalization.* BMC Res Notes, 2018. **11**(1): p. 156.
22. White, A.M., et al., *ProMAT: protein microarray analysis tool.* Bioinformatics, 2006. **22**(10): p. 1278-9.
23. Almamy, A., et al., *Crossreactivity of an Antiserum Directed to the Gram-Negative Bacterium Neisseria gonorrhoeae with the SNARE-Complex Protein Snap23 Correlates to Impaired Exocytosis in SH-SY5Y Cells.* J Mol Neurosci, 2017. **62**(2): p. 163-180.
24. Almamy, A., et al., *Interactions of antisera to different Chlamydia and Chlamydophila species with the ribosomal protein RPS27a correlate with impaired protein synthesis in a human choroid plexus papilloma cell line.* Immunol Res, 2017. **65**(6): p. 1110-1123.
25. Ludwig, N., et al., *Improving seroreactivity-based detection of glioma.* Neoplasia, 2009. **11**(12): p. 1383-9.
26. Becker, A., et al., *Myasthenia gravis: analysis of serum autoantibody reactivities to 1827 potential human autoantigens by protein macroarrays.* PLoS One, 2013. **8**(3): p. e58095.
27. Sahlstrom, P., et al., *Different Hierarchies of Anti-Modified Protein Autoantibody Reactivities in Rheumatoid Arthritis.* Arthritis Rheumatol, 2020. **72**(10): p. 1643-1657.
28. Wobst, H., et al., *UCHL1 regulates ubiquitination and recycling of the neural cell adhesion molecule NCAM.* FEBS J, 2012. **279**(23): p. 4398-409.