Advances in the development of human protein microarrays

Jessica G. Duarte & Jonathan M. Blackburn

To cite this article: Jessica G. Duarte & Jonathan M. Blackburn (2017) Advances in the development of human protein microarrays, Expert Review of Proteomics, 14:7, 627-641, DOI: 10.1080/14789450.2017.1347042

To link to this article: http://dx.doi.org/10.1080/14789450.2017.1347042

Accepted author version posted online: 23 Jun 2017.
Published online: 04 Jul 2017.
Advances in the development of human protein microarrays

Jessica G. Duarte and Jonathan M. Blackburn

ABSTRACT

Introduction: High-content protein microarrays in principle enable the functional interrogation of the human proteome in a broad range of applications, including biomarker discovery, profiling of immune responses, identification of enzyme substrates, and quantifying protein-small molecule, protein-protein and protein-DNA/RNA interactions. As with other microarrays, the underlying proteomic platforms are under active technological development and a range of different protein microarrays are now commercially available. However, deciphering the differences between these platforms to identify the most suitable protein microarray for the specific research question is not always straightforward.

Areas covered: This review provides an overview of the technological basis, applications and limitations of some of the most commonly used full-length, recombinant protein and protein fragment microarray platforms, including ProtoArray Human Protein Microarrays, HuProt Human Proteome Microarrays, Human Protein Atlas Protein Fragment Arrays, Nucleic Acid Programmable Arrays and Immunome Protein Arrays.

Expert commentary: The choice of appropriate protein microarray platform depends on the specific biological application in hand, with both more focused, lower density and higher density arrays having distinct advantages. Full-length protein arrays offer advantages in biomarker discovery profiling applications, although care is required in ensuring that the protein production and array fabrication methodology is compatible with the required downstream functionality.

1. Introduction

Since the introduction of DNA microarrays in the mid-1990s, a wide range of analogous microarray technologies have been spawned that enable a multitude of applications across the biological and biomedical sciences and which encompass genomic, transcriptomic, proteomic, glycomic, epigenomic, and drug discovery research areas, among others. At their simplest, microarrays can be considered as lab-on-chip tools consisting of multiple different biomolecules that are immobilized at spatially addressable locations on a surface and which can function as discrete probes in downstream, highly miniaturized, multiplexed assays. The parallel interaction of such immobilized probes with analytes contained in complex sample solutions results in the generation of vast amounts of quantitative, functional and interaction-based information for analysis [1,2].

The underlying microarray technologies themselves remain under active development in several areas, including diversification of content, improvement in surfaces and broadening of application areas. For example, in the genomic field, DNA microarray technology has been revolutionized over the years by dramatically expanding the number of probes on the arrays while concomitantly reducing the feature sizes, improving the underlying surfaces, and developing many new application areas, including inter alia global analysis of gene expression, copy number variation, splice variation, polymorphic variation, DNA methylation, and transcription factor binding sites.

In the proteomics field, microarray technologies today include full-length protein arrays, protein fragment arrays, peptide arrays, antibody arrays, reverse-phase arrays and tissue arrays. In this Review, we provide an overview of recent technological advances in the development of protein microarrays, with a specific focus on full-length protein and protein fragment microarrays, discussing the technological basis, applications and limitations of some of the most commonly used protein microarray platforms. Technological advances in the antibody microarray field have been reviewed recently elsewhere [3–7] so are not dealt with again here.

Early research in the protein microarray field focused initially on simple technical demonstrations that instrumentation and surfaces adapted directly from the DNA microarray field could also be used to fabricate protein microarrays: commercially-available arrayers were used to print nanolitre volumes of different purified proteins onto chemically derivatized glass slides, resulting in high density, spatially defined patterns; discrete locations on those prototype microarrays were identified through on-chip protein interaction assays, using fluorescently labeled, known binding partners; and assays were read out using DNA microarray scanners [8]. From those
relatively modest beginnings, protein microarrays have developed progressively into valuable proteomic tools that today can contain tens of thousands of different proteins, including entire proteomes in some cases, and which are capable of addressing numerous different biological questions, with potential applications in biomarker discovery as well as in the quantitative analyses of protein function and interactions. However, realization of the true potential of modern protein microarray technologies relies on the effective integration and application of various components of the overall system, including high throughput protein expression and purification, protein immobilization, assay development, signal detection, data processing and data analysis [9]. Classic limiting factors to the wider uptake of protein microarrays by the community have included the availability of protein content in a form suitable for arraying, the development of surface chemistries to preserve the folded structure and function of proteins on immobilization, and the availability of assays for all the arrayed proteins; these factors are discussed briefly in turn below.

1.1. Content generation

One of the early decisions to be made in the protein microarray field is how the content will be generated. One obvious approach is to separate many different proteins directly from native sources, which at first sight has the advantage that the native proteins come complete with relevant post-translational modifications and potentially also with interacting partners. However, purification of a single native protein or complex from a biological sample can be a lengthy, highly optimized process. By comparison, multi-dimensional fractionation strategies to produce the content for native protein microarrays are not optimized for each individual protein in the sample, so instead yield simplified mixtures of unknown composition and unknown concentration that require further downstream deconvolution. Furthermore, reproducibility of protein purification and activity from native sources remains an issue, even for abundant native proteins such as plant lectins. Thus, despite the cost and labor-intensive nature of high throughput cloning and recombinant expression, the major protein microarray platforms available today focus largely on recombinant sources, with proteins expressed heterologously in either simpler prokaryotic hosts (usually *Escherichia coli*), in yeast, insect or mammalian cells, or in cell-free transcription-translation systems.

Each heterologous expression system has its own advantages: *E. coli* is quick to culture and easy to lyse in high throughput and can give good yields of recombinant protein, but suffers from the disadvantages that prokaryotes and eukaryotes use different co- or post-translational protein folding mechanisms and that it lacks eukaryotic-like post-translational modification (PTM) machinery, resulting in lower success rates for high throughput expression of functional eukaryotic proteins. By comparison, yeast expression systems are better suited to the heterologous expression of eukaryotic proteins with simple eukaryotic PTMs, but suffer from difficulties of cell lysis without induction of endogenous proteases, which complicates downstream high throughput purification strategies. Insect cell expression systems have therefore found favor with several groups since, once transfected with a suitable baculoviral vector, expression of many different eukaryotic proteins with more complex PTMs can be achieved in parallel in a few days, albeit with typically lower yields, and downstream lysis methods are much milder than for yeast [10]. By contrast, mammalian expression systems are less simple to establish and maintain in high throughput, while cell-free transcription-translation systems can be good sources of polypeptide, but often require optimization for expression of individual folded, functional proteins.

Following expression, high throughput, parallel purification of recombinant proteins to near-homogeneity is also not always straightforward, since in many cases, short peptide tags (e.g. the hexa-His tag) can be occluded, necessitating purification under denaturing conditions [11]. The protein microarray field has thus in general gravitated now toward use of larger protein domains as affinity tags (e.g. the glutathione S-transferase (GST) tag), as well as to the use of automated purification systems. Some groups, though, have adopted alternative purification strategies, combining specific protein affinity tags with specialized array surfaces to enable single step, *in situ* protein purification and surface immobilization [12–17], thus circumventing the need for laborious high throughput pre-purification of proteins prior to array fabrication.

1.2. Array fabrication

The construction of a high-density protein microarray, involving the immobilization of diverse protein content, requires numerous factors to be considered in order to assure an optimal microarray design [2].

High content protein microarray fabrication methods typically use either contact printing or piezoelectric deposition methods, with instruments adapted directly from the DNA microarray field. Typically, piezoelectric printing methods give excellent, reproducible control over spot morphologies and spot sizes, but are prone to blockage of nozzles if the samples are too viscous or contain particulate matter. Contact printing with either solid or split pins is more robust, but can result in greater spot-to-spot variability due to the subtle variation in the machining of individual pins. Irrespective of the exact printing method used, care needs to be taken in all fabrication processes to avoid carry-over between different sample analytes deposited with the same pin/nozzle, so inclusion of suitable wash steps and quality control checks is therefore essential [18].

Consideration also needs to be applied to the choice of buffer from which the proteins are printed, since the surface tension of the printed droplets affects evaporation rates as well as spot spreading, so influences spot morphology, spot size and homogeneity of the immobilized protein within each spot. Spot sizes can be altered based on the specific protein microarray application and content density required and are typically in the range of 50–450 μm – considerably larger than the feature sizes on current commercial DNA microarrays.
Spot-to-spot spacing also needs to be large enough to minimize spot running and cross-contamination during fabrication. The use of duplicate, triplicate or quadruplicate replicates per probe is commonplace on protein microarrays, since this assists in assay validation and in determination of coefficients of variation within and across replica arrays. Appropriate positive and negative controls should also be included in array designs where possible.

Important choices also need to be made regarding the surface chemistries on to which the arrays are fabricated, in order to ensure that sufficient protein densities per spot are achieved, as well as to control orientation and folding of arrayed target proteins when required. The simplest fabrication approaches make use of either chemically-reactive glass surfaces (typically either aldehyde-, N-hydroxy succinimide- or epoxide-activated) or of nitrocellulose-coated glass [8,19]; in both cases, immobilization of essentially all proteins can be readily achieved in parallel, but the binding to such surfaces is uncontrolled, through random covalent coupling or non-specific physisorption, respectively and can lead to unfolding and loss of activity. Other approaches therefore aim to achieve controlled, oriented immobilization of different recombinant proteins onto the surface via a common affinity tag [2,12,14,17,20,21], which has theoretical advantages, particularly if downstream, quantitative, on-array functional assays are required [20].

Post-microarraying, typically some form of blocking agent is applied to the surrounding surface in order to abrogate non-specific binding of analyte molecules to the underlying surface during downstream assays. Here, typical methods simply adapt blocking agents commonly used in enzyme-linked immunosorbtent assays (ELISA), such as bovine serum albumin or milk proteins. In the absence of specific control over on-array protein stability, many groups simply print protein microarrays at point of use, potentially creating unwanted batch effects in downstream assays. However, more sophisticated protein microarray architectures make use of the well-known and superior properties of polyethylene glycol polymers, which block non-specific macromolecule binding to the surface [21–23] while at the same time creating a hydrated, hydrogel-like environment that helps stabilize the folded state and functionality of the immobilized proteins for downstream assay over periods of many months. Such stability of arrayed proteins post-immobilization is a key factor in enabling the use of protein microarrays as functional and quantitative analytical tools [18].

### 1.3. Assay development

Once high-density protein microarrays with diverse protein content have been constructed, the next hurdle to overcome is the development of assays that work for all the arrayed proteins. In contrast to DNA microarrays, this has remained a major bottleneck in the protein microarray field since the different arrayed proteins typically have very different physical and biochemical properties, as well as widely differing affinities for their interacting analyte molecules. Numerous papers have reported functional assays carried out on-array on smaller, focused sets of proteins, for example quantitative protein-DNA binding assays across sets of transcription factors [20], kinase substrate identification assays [20,24] and protein–protein interaction mapping assays [25]. However, as the protein content on the microarrays becomes more diverse, the range of array-based assays that can be configured for every microarrayed protein diminishes due to the lack of commonality in structure or function of the arrayed proteins and to the still-limited methods for detecting the results of microarray-based assays. The first protein microarray application to have taken hold widely therefore is serological screening. Here, the interaction between target analytes (e.g. antigen-specific antibodies) contained in a sample solution (e.g. human serum or plasma) with arrayed proteins (antigens) can be readily detected since, while the target antigens themselves are diverse, the constant regions of the antibody analytes enables readout of binding interactions via, for example, a fluorescently-labeled secondary antibody [2,26–28].

As with ELISAs, in such serological assays, it is critical that cross-reactivity of the secondary antibody is eliminated, to minimize the possibility of false positives [18]. However, for quantitative serological analysis, it is equally important that the measured signal in the assay does not inadvertently depend on the affinity of the labeled secondary antibody for the captured primary antibody, which argues that the secondary antibody needs to be used at a concentration that ensures saturation binding [21]. Thus, considerable optimization of such serological assays is still required, despite their apparent simplicity.

Detection limits vary between different microarray platforms, depending in part on the choice of detector, but the most widely used microarrays typically offer high sensitivity in fluorescence-based assays that are read-out on a confocal DNA microarray scanner, thereby in principle enabling both qualitative and quantitative data to be produced. As with classic ligand binding assays, the limit of detection in many such assays is thus fundamentally dominated by physical chemistry considerations, based on the combination of the intrinsic binding affinity of the capture probe for the target analyte and the concentration of that analyte in solution [29]. The limit of detection in microarray assays is also controlled by the signal to noise ratio in the assay, with the lower background binding afforded by, for example, polyethylene glycol-coated surfaces enabling a lower limit of detection of true foreground signal [21] compared to more traditional blocking agents that show higher non-specific background binding.

### 1.4. Data processing and analysis

Protein microarrays are usually high-content, allowing for the high throughput analysis of numerous target antigens in parallel. In addition, protein microarrays can be printed with multiple replica sub-arrays on each slide (e.g. in 4-plex or 16-plex formats), with each replica array on a given slide being assayable independently with either different samples or different concentrations of a given analyte, but under otherwise uniform assay conditions [2,20]. Protein microarrays can thus yield large amounts of raw data, which require dedicated processing and analysis tools to enable the correct biological interpretation of the results. It is important to note here that
while many processing and analysis tools exist from the DNA microarray field, typically these cannot be used readily for the analysis of protein microarray datasets because the underlying statistical assumptions differ between the microarray experiments. For example, in a gene expression study on a DNA microarray, it is typically assumed that a roughly equal number of genes will be up- and down-regulated as a result of the test condition, so the mean expression across all genes, as well as the distribution of expression values, should remain the same and quantile normalization can therefore be applied to the raw datasets. However, in a protein microarray experiment, typically fewer elements on the array will produce a true positive signal in the assay and these might differ substantially in identity and magnitude between test and control assays for genuine biological reasons, in which case both the mean and the distributions will be altered and quantile normalization therefore should not be applied [2,30].

Several tools are currently available that have been specifically developed for protein microarray datasets, offering bioinformatic and statistical analysis methods that are suitable for this type of data [31]; some of these are described in more detail in subsequent sections of this review. As in the DNA microarray and proteomics fields, there is no single standard data processing or analysis pipeline used by all protein microarray users, but rather each research group typically adopts their own specialized methods, which creates problems when trying to compare reported protein microarray datasets between laboratories. Given that array content and surface chemistries differ between protein microarray platforms, if a standardized set of reporting criteria were in place (similar, for example to the MIAME standards in the DNA microarray field), datasets produced by different labs could become complimentary to one another within a common disease setting. With the ambition of enabling a more generically useful data processing pipeline for protein microarray datasets, a composite and comprehensive bioinformatic tool has recently been developed for raw data processing which includes methods such as neighborhood background correction and subtraction, noise and CV thresholds, data point and array filtering, pin-topin and array-to-array normalization, and data consolidation [2].

A further consideration in the field is how to objectively determine the threshold for true signal. The simplest approach here relies solely on subtraction of the signal from the surrounding background, but with the availability of pixelated data, it is straightforward to calculate the standard deviation (SD) of the surrounding background signal and to then set a classic noise threshold for true signal of 2SD from the mean background pixel intensity [2]. However, the weakness in this approach is that the surrounding area has a different biochemical composition to the foreground probe area, so the non-specific binding of analytes to the foreground and background areas is not necessarily the same. A more nuanced approach therefore relies on negative control data for each arrayed protein to define a probe- and assay-specific threshold for true signal.

Following the use of appropriate data processing tools, the biological interpretation of the resulting extracted datasets can be achieved by way of data clustering and fold-change based statistical methods, among others, depending on the research question in place. Bioinformatic methods for such downstream analyses are typically adapted from the genomic field. Since there is no consistency between research groups in the choice of statistical approach to data analysis in either the protein or DNA microarray fields, it remains of critical importance that the validity of the statistical analysis is considered up front. Importantly, complications can arise if genomic analysis methods are applied to protein microarray datasets without prior consideration of the underlying data structure. For example, hierarchical cluster analysis of an unfiltered protein microarray dataset that genuinely contains large numbers of zero values and small numbers of positive values will be considerably less informative than hierarchical cluster analysis of the same dataset that was pre-filtered to focus only on the most commonly identified proteins [32]; this can be likened to a penetrance-based analysis in the genomics field. Furthermore, as with all ‘omics experiments, it is important to consider the sample size and number of technical replicates per sample that are required to achieve the statistical power required for the study.

Protein microarrays currently enable the interrogation of an ever larger subset of the human proteome and of the entire proteome of specific microorganisms (e.g. E. coli, yeast, Mycobacterium tuberculosis, and human papilloma virus). Such protein microarrays are now finding utility across a broad number of different biological applications, including biomarker discovery in cancers, autoimmune diseases, neurological diseases and inflammatory diseases [33]. In the following sections of this Review, we will discuss some of the most commonly used planar, recombinant protein microarray platforms, highlighting the key differences and uses of each of these.

2. ProtoArray human protein microarray

The progenitor of the ProtoArray human protein microarray (Thermo Fisher Scientific) was a yeast proteome microarray, originally developed by the Snyder laboratory at Yale containing 5800 full-length, purified target yeast proteins that were used to screen for diverse biochemical activities [34]. These N-terminal GST-His-tagged proteins were produced recombinantly from high-quality yeast ORF clones and expressed in Saccharomyces cerevisiae to ensure correct protein folding. After high throughput purification, proteins were printed onto aldehyde-treated or nickel-coated slides, with immobilization being via amine or His-tag attachment respectively, and detected using fluorescently-labeled anti-GST antibodies; immobilization via the His-tag was found to yield superior signals. Yeast is often referred to as the model host and studying the yeast proteome has enabled the understanding of many essential cellular processes of relevance to higher eukaryotes [35]. The yeast proteome microarray has been used across numerous functional applications, including protein-protein interactions, phospholipid interactions and small molecule interactions, as well as verification of antibody targets in antibody target assays, and the identification of substrates for solution phase enzymes [36].
The ProtoArray human protein microarray today consists of 9483 unique full-length and functional human proteins and appropriate controls. This content includes disease-relevant proteins across several distinct protein classes, including protein kinases, transcription factors, membrane proteins, nuclear proteins and secreted proteins, as well as signal transduction, cell communication, metabolism, cell death, and protease activities. These human proteins are produced from sequenced-verified clones (Invitrogen Ultimate ORF collection) and are expressed as N-terminal GST fusion proteins using a baculovirus expression system (Gateway entry vector pENTR221), with high throughput purification carried out under non-denaturing conditions. Purified proteins and controls are subsequently printed in duplicate on thin-film nitrocellulose-coated glass slides at 4°C. The ProtoArray human protein microarray can then be incubated with assay sample, washed and probed with fluorescently labeled detection reagents, prior to visualization on any open architecture fluorescence microarray scanner.

The resulting data can be analyzed using the free ProtoArray Prospector software which features algorithms for analysis of the various supported applications, as well as for data normalization, and provides a report output. This software includes signal normalization using a z-score method, which indicates how an individual data point differs from the population mean in units of standard deviation:

$$z\text{-score} = \frac{(X - \mu)}{\sigma},$$

where $X$ is the signal value from a protein feature, $\mu$ is the mean signal for all protein features, and $\sigma$ is the signal sample standard deviation for all protein features. Noise thresholds are then set using the signal-to-noise ratio z-factor method, which measures the dynamic range and the variation between noise and signals:

$$z\text{-factor} = \frac{3(\sigma_n + \sigma_\mu)}{|\mu - \mu_n|},$$

where $\sigma_n$ is the signal sample standard deviation for the protein features, $\sigma_\mu$ is the signal sample standard deviation for the negative control features, $\mu$ is the mean signal for the protein features and $\mu_n$ is the mean signal for the negative control features. A z-factor > 0.5 indicates a signal greater than twofold above the array background. Outlier identification is carried out using the Chebyshev’s inequality (CI) p-value method, which compares the signal from each protein feature to the distribution of negative controls:

$$CIp - value = \begin{cases} 1 & X \leq \mu_n + \sigma_n \\ \frac{\sigma_n}{\mu - \mu_n} & X > \mu_n + \sigma_n \end{cases},$$

where $X$ is the signal value from a protein feature, $\mu_n$ is the mean signal for the negative control features and $\sigma_n$ is the signal sample standard deviation for the negative control features. This algorithm calculates the p-value for the null hypothesis that a given signal belongs to the negative control features distribution.

Profiling applications of the ProtoArray human protein microarray promoted by the manufacturer include disease-specific biomarker discovery using immune response profiling, target discovery and validation using enzyme substrate profiling, target identification or selectivity using small molecule profiling, mapping biochemical pathways using protein-protein interaction profiling, and therapeutic antibody development using antibody specificity profiling. To date, the majority of publications based on use of the ProtoArray platform have focused on the first of these areas: qualitative profiling of antibody responses in disease.

Within the cancer biomarker discovery field, serum antibody response profiling has been carried out by numerous groups using the ProtoArray platform on a broad spectrum of cancers, including bladder cancer [37], breast cancer [38], chronic lymphocytic leukemia [39], colorectal cancer [40], lung cancer [41], myelodysplastic syndromes [42], ovarian cancer [43,44], and pancreatic cancer [45].

Additional disease-specific biomarker discovery studies have been performed for autoimmune diseases (rheumatoid arthritis [46], Sjögren’s syndrome [47], type 1 diabetes [48], Meniere’s disease [49]), autoimmune polyendocrine syndrome type 1 [50]), inflammatory diseases (asthma [51], inflammatory bowel disease [52]), neurological diseases (Alzheimer’s disease [53], Parkinson’s disease [54], multiple sclerosis [55], amyotrophic lateral sclerosis [56]), organ transplants (kidney allografts [57], kidney transplant [58], chronic transplant injury [59]), and chronic diseases (chronic renal disease [60]).

However, despite the seemingly broad applicability of these microarrays, a number of issues arise that potentially complicate deduction of biological significance from the ProtoArray data. Porous nitrocellulose surfaces, including the thin-film nitrocellulose slides used in fabrication of the ProtoArray, give higher protein binding capacities than planar surfaces, but typically suffer from higher intrinsic background fluorescence, as well as from higher non-specific binding to the surface [61] and reduced stability and reproducibility [62]. A number of different approaches have been explored to minimize background fluorescence issues with nitrocellulose surfaces, the aim being to increase signal-to-noise ratios in assays, including reducing the thickness of the nitrocellulose coating [63] and use of black nitrocellulose membranes [64]. These modifications may though negatively affect protein binding capacities and signal homogeneity [65]. Some newer microarray scanners can also read reporter probes with emission in the near-infrared region (700–800 nm), away from the commonly used shorter wavelength regions of the spectrum (e.g. 532 and 635 nm) where autofluorescence is particularly high due to light scattering within the 3D nitrocellulose matrix, but even this does not directly resolve issues with higher non-specific binding.

More importantly though, while the individual recombinant proteins on the ProtoArray platform are expressed in insect cells and purified under native conditions, the nitrocellulose microarray surfaces provide a relatively denaturing environment to non-immunoglobulin proteins. Moreover, proteins can also be at least partially stripped away from the nitrocellulose surface during assay [66], potentially leading to some lateral protein diffusion on the array surfaces. Any unfolding of individual proteins on the nitrocellulose surface will disrupt
discontinuous epitopes, potentially leading to false negative results in serological profiling assays, and may also result in false positives, through non-specific antibody binding to exposed hydrophobic epitopes that areordinarily buried in the native structures of the arrayed proteins (N.B.: this phenomenon is frequently seen in western blots of denaturing polyacrylamide gels, with both monoclonal and polyclonal antibody reagents). Furthermore, the ProtoArray Prospector software has some scope for improvement, particularly when dealing with large data sets, through the implementation of more appropriate normalization methods and other desirable features that have been discussed elsewhere in detail [2,31,67].

3. HuProt human proteome microarray

The HuProt Human Proteome Microarray (CDI Laboratories) was originally developed by the Zhu laboratory at Johns Hopkins University and contained 16,368 unique full-length human proteins, representing 12,586 protein-coding genes [68]. If one adopts the simplistic ‘one gene = one protein’ paradigm (i.e. ignoring splice variants and PTMs), this represents ~60% of the human proteome. Similar to the yeast proteome microarray described above, in the HuProt platform, human proteins are expressed recombinantly in S. cerevisiae as N-terminal GST-His-tagged fusion proteins, using a pEGH-A yeast expression vector [69]. After high throughput affinity purification using the GST tag, proteins and controls are printed in duplicate onto glass slides, and detected for quality control purposes using an anti-GST antibody. Among a wide range of potential applications, this high density human protein microarray has found utility, for example, in the low-cost identification and validation of novel antibodies.

The current high-content v3.0 HuProt array contains >19,000 full-length human proteins, corresponding to >15,000 unique protein-coding genes. These proteins belong to multiple distinct categories, including kinases, cell death, membrane proteins, nuclear proteins, metabolism, signal transduction, secreted proteins, and transcription factors. The HuProt arrays can be printed on a variety of different glass slides, including FAST, FullMoon, SuperEpoxy, SuperAldehyde, SuperNHS, NiNTA, and PATH. For antibody profiling applications, after blocking, arrayed slides are incubated with the biological sample and bound IgG antibodies are detected using a fluorescently-labeled human IgG secondary antibody. Assayed slides can then be scanned with any open architecture microarray scanner, using the GenePix array list (.gal file) from the arrayers that describes the location of each protein on the array. The resulting data can then be analyzed using various image acquisition and analysis software such as GenePix, ArrayPro, or others. In the absence of a standardized downstream data analysis package or approach, data processing, normalization and analysis methods have varied considerably between studies based on the HuProt arrays, including use of methods such as z-scores, Fisher’s exact test on positive incidence and rate, signal value cut-offs, t-tests, quantile normalization, outlier detection, false discovery rate detection and ROC curves; these methods are not all equally statistically valid for protein microarrays though, which represents a confounding factor for researchers using this technology without specialized knowledge in the field.

Antibody profiling with the HuProt array – also referred to as biomarker profiling – has been carried out across numerous disease settings, including primary biliary cirrhosis [70], neuropsychiatric lupus [71], glioma [72], Behcet Disease [73], secondary progressive multiple sclerosis [74], and gastric cancer [75]. Additional uses of this array include investigating antibody specificity, small molecule profiling, protein–protein interactions, DNA/RNA binding and enzyme substrate identification.

Using an essentially identical approach to protein production, purification, and immobilization to that described above, other related arrays have also recently been described on the same basic platform, for example an E. coli proteome microarray [76–78] and a M. tuberculosis proteome microarray [79], but have not yet gained widespread utilization beyond the academic groups who developed these arrays.

A potential weakness of the HuProt system however lies in the reliance on expression of human proteins in a lower eukaryote, in which post-lysis proteolysis is a known problem and in which only simpler post-translational modifications will occur. Furthermore, as with the ProtoArrays, another potential limitation of the HuProt system lies in the surface chemistry used for array fabrication: non-specific immobilization of the arrayed proteins onto the glass surface, whether by physisorption or covalent binding, can result in unfolding that confounds downstream data interpretation, as discussed earlier.

4. Human Protein Atlas protein fragment array

The Human Protein Atlas is an open-access interactive database containing extensive images and data derived from antibody-based proteomics and transcriptomics experimentation [80–83]. The Human Protein Atlas uses both in-house and commercially-generated antibodies for immunohistochemistry on tissue microarrays (to document the distribution of protein expression in normal and cancer tissues) and for immunofluorescence staining on cell lines (to document the spatial distribution at a subcellular level). The resultant database includes the majority of human protein-coding genes, with information about the expression and localization of proteins, and is subdivided into three different atlases: the Tissue Atlas provides pathology-based annotated protein expression levels and immunohistochemistry stained images obtained across 76 different cell types, corresponding to 44 different normal tissue types, as well as RNA deep sequencing-based RNA expression data obtained across 37 different normal tissue types; the Cancer Atlas provides pathology-based annotated protein expression levels and immunohistochemistry stained images obtained across 216 different cancer samples, corresponding to 20 different cancer types with basic demographic and tumor phenotype information; and the Cell Atlas provides antibody profiling-based annotated protein expression levels and immunofluorescence confocal microscopy stained images obtained across 56 different cell lines, corresponding to 32 different organelles and fine cellular structures, as well as RNA deep sequencing-based RNA expression data.
In order to enable the large-scale generation of high specificity antibodies to human proteins, the Human Protein Atlas consortium employs a bioinformatic strategy to identify ‘protei

teptide signature tags’ (PrESTs) for each human protein \[84,85\], representing the peptide regions of each protein which show the lowest similarity to any other human protein. These PrESTs are then produced and used in high throughput immunization of animals to elicit the cognate antibodies. This bioinformatic approach combines the PRESTige web-based tool and a whole-genome analysis to avoid 50-amino acid regions with more than 60% sequence identity to other protein domains, 10-amino acid regions with more than 8 identical amino-acids, transmembrane regions, signaling peptides and restriction enzyme recognition sites commonly used in cloning methods.

The resulting PrEST fragments are typically 80-amino acids long, ranging between 16 and 202-amino acids, and are considered unique representations of the corresponding proteins; these fragments are expressed as N-terminal albumin binding protein (ABP)-His-tagged fusion using the pAff8c expression vector in *E. coli* Rosetta DE3 to increase protein yields for immunization \[86,87\].

These same PrEST fragments have subsequently found additional utility as the content for high-density microarrays, in both planar and bead-based suspension microarray formats \[81\]. The resultant Human Protein Atlas Protein Fragment Arrays developed by the Nilsson laboratory at KTH (accessible via the Autoimmunity Profiling Facility, ScilifeLab, Stockholm) and currently include 42,698 unique purified protein fragments, derived from 19,497 human protein coding genes \[74\]; using the SwissProt ‘one-gene = one-protein’ paradigm, 94% of the human proteome is thus represented by at least one fragment on these arrays.

For the production of Human Protein Atlas Protein Fragment Arrays, an optimized high-throughput protein production workflow is used for the large scale expression, affinity purification and verification (by mass spectrometry) of the recombinant proteins. Purified protein fragments are spotted onto epoxide-coated glass slides using a non-contact microarray printer and the residual surfaces then blocked \[74\] in a similar manner to HuProt arrays. These high density microarrays have found utility in, for example, the validation of newly generated Human Protein Atlas antibodies and in autoantibody profiling of serum and plasma.

Autoantibody profiling with such protein fragment arrays has been carried out across different diseases, including multiple sclerosis, sarcoidosis, and osteoarthritis, often as a screening or discovery phase, followed by validation on bead-based suspension arrays \[74,88–91\]. As with other protein microarray platforms, the data analysis pipelines and statistical approaches used varies between studies, although the majority convert the resulting fluorescent intensities to a binary variable by defining a sample-specific threshold (signal intensities > median + n × median absolute deviation or standard deviation), and apply a comparative false discovery rate-corrected statistical test to compare sample groups and identify significant differences.

At first sight, the Human Protein Atlas Protein Fragment Arrays offer an apparent advantage over other protein array platforms, as a result of the high proportion of the human proteome that is represented on these arrays. However, it is important to recognize that there are typically only two PrEST tags per human protein on the arrays and that any one tag represents only a fraction of all possible epitopes on any given protein. Furthermore, discontinuous epitopes and post-translationally modified epitopes will also be largely absent on these protein fragment arrays, so if the coverage was represented as a fraction of all possible human epitopes, it would be significantly smaller than the claimed 94%. Moreover, the use of protein fragments (which presumably in general have little tertiary structure), rather than full-length, folded proteins, restricts the range of potential assays on this platform largely to profiling of antibody binding to a repertoire of linear epitopes \[74\]; in this regard, the Human Protein Atlas Protein Fragment Arrays can therefore be considered closer conceptually to high density peptide arrays than to full length protein arrays.

### 5. Nucleic Acid Programmable Protein Array (NAPPA)

An innovative, alternative approach to the production of protein microarrays was first reported by the Taussig group (Babraham Institute, UK) and developed more fully by the LaBaer group (Harvard), dispensing with off-line, high throughput protein expression and purification altogether. Although the precise components differ, essentially in the Protein *In Situ* Array (PISA \[12\]) and Nucleic Acid Programmable Protein Array (NAPPA \[14\]) methods, cDNAs encoding an in-frame fusion to a tag are printed (but not immobilized) into a spatially defined array, together with a capture reagent that is specific for the tag (the capture agent is itself immobilized on the array surface). The encoded, tagged proteins are then expressed *in situ* by overlaying coupled cell-free transcription-translation mixes and each tagged protein then binds rapidly to its cognate, immobilized capture reagent to form the protein microarray. More recent technological developments of this *in situ* array fabrication approach have focused on the choice of tag:

In the PISA method, a C-terminal double His\(_6\)-tag was used, but the relatively low binding affinity of a tag-surface combination really designed for protein purification results in relatively high spontaneous off-rates from Ni-NTA surfaces and any subsequent re-binding events would begin to scramble the spatial locations on the array;

In the original NAPPA method, a C-terminal GST tag was used, combined with an immobilized polyclonal anti-GST antibody to provide a more stable linkage. However, the antibody itself was adsorbed non-specifically through ionic interactions to an amino-silanized glass surface, so spontaneous desorption and re-binding (equating to time-dependent, lateral diffusion of the arrayed proteins) was again likely an issue;

In the most recent manifestation of the NAPPA method, a C-terminal HaloTag \[17,92\] is used to provide covalent binding of the tagged protein to the surface via a chloroalkane ligand that is co-arrayed and is itself cross-linked on to an aminosilane glass slide. The authors note that this approach results in rapid capture and up to four times higher densities of expressed fusion proteins per spot than the original NAPPA.
method, resulting in higher signal-to-noise ratios; importantly, they also note that there is minimal lateral protein diffusion as a result of the covalent immobilization method.

Human protein NAPPA arrays are today available from the Protein Array Core of the BioDesign Institute (Arizona) and contain up to 11,500 human ORF expression products printed in a variety of formats. NAPPA arrays have also been produced for other organisms, including Vibrio cholera, M. tuberculosis, and Ornithodoros moubata, as well as Arabidopsis protein arrays comprising a total of 12,000 ORFs (each containing 4600 ORFs printed in duplicate [17,93]). NAPPA arrays have been used in a variety of serological studies, including in diabetes [94,95], juvenile arthritis [96], ankylosing spondylitis [97], breast cancer [98], ovarian cancer [99] and varicella zoster virus infections [100], as well as in mapping transcription factor interactome networks in Arabidopsis [17] and detecting post-translational modifications such as AMPylation [101] and chemical acetylation [102].

The simplicity of the NAPPA array fabrication is attractive, not least since assays can be carried out on freshly expressed proteins. However, it is less clear whether the NAPPA arrays will prove to be compatible with a wider variety of assays for a number of reasons. In particular, cell-free transcription-translation systems have classically proved to be a good way to produce polypeptide in increasingly large scale, but can require considerable optimization of conditions to produce any given protein in a folded, functional form, perhaps because there is limited control over redox conditions during expression and folding due to the loss of compartmentalization [16]; this is likely to be particularly problematic for more complex proteins with multiple disulphide bonds and will be exacerbated by the fact that all proteins are by necessity expressed under a single set of in vitro conditions for the NAPPA arrays. Furthermore, it seems likely that post-translational modifications on eukaryotic proteins produced by cell-free transcription-translation will be simpler, even if eukaryotic cell-free extracts are used, again due to compartmentalization issues. As with the ProtoArray and HuProt systems, only simple surface chemistry has been used to date for NAPPA arrays (aminosilane glass slides) but concerns related to slow on-array unfolding are likely to be less of an issue here because the NAPPA arrays are typically produced to order at point of use.

6. Immunome protein array

The Immunome protein array platform (Sengenics) was originally developed by the Blackburn laboratory at the University of Cambridge, based on the concept of lower density microarrays comprising focused families of full-length, folded, functional proteins and designed from the bottom up to enable the quantitative, high throughput, systems-oriented analysis of a wide range of different protein functions. For example, different protein function microarrays from the Blackburn laboratory included: ~330 human kinases, representing roughly half the human kineome [15]; ~100 cancer-testis antigens, representing ~80% of that protein family [2,21]; ~300 human transcription factors; and ~50 clinically-relevant polymorphic variants of p53 [20]. Published assays on this platform include: quantitation of protein-DNA and protein-drug binding affinities and selectivity across families of structurally-related proteins [15,20]; immune response profiling [21,103–105]; on-array post-translational modifications [20]; and measurement of protein–protein interactions [20].

Currently, the Immunome protein array contains 1636 full-length, correctly folded and functional human proteins plus controls, representing major protein classes involved in the immune response, such as cancer antigens, interleukins, cytokines, protein kinases, transcription factors and signaling proteins. Each protein is printed in quadruplicate across a single array to enable more rigorous statistical analysis of quantitative data downstream.

The proteins on the Immunome protein array are expressed from baculoviral vectors in insect cells as fusions to a C-terminal biotin carboxyl carrier protein (BCCP) tag, which is itself a compact, 80 amino acid, all-beta domain of the E. coli acetyl CoA carboxylase [106]. Expression of human proteins in insect cells is well known to have no particular size limitation and, since intrinsic folding mechanisms are shared with higher eukaryotes, both folding and post-translational modifications are considered to be mammalian like [107,108].

The BCCP tag is biotinylated in vivo on a single lysine residue (K122, based on AccB numbering) by host cell biotin ligases in E. coli, yeast, insect cells and mammalian cells, without the need to co-express the E. coli biotin ligase. The BCCP tag is however only biotinylated in vivo if the tag itself folds into its correct tertiary structure and, unlike the in vitro evolved AviTag, the linear biotinylation motif in BCCP is not a substrate for the biotin ligases [109]. In common with other small, autonomously folding bacterial proteins (e.g. glutathione S transferase, thioredoxin), an N-terminal BCCP tag can increase expression and solubility of a C-terminal fusion partner [20]. Moreover, if used as a C-terminal tag, post-translational modification (i.e. biotinylation) of BCCP has been found to be a useful reporter for the folded state of fusion partners [15,21,110] in a manner similar to other known folding markers, such as the green fluorescent protein (GFP) [111,112]. In the case of GFP, post-translational formation of the fluorophore does not occur if the upstream fusion partner itself mis-folds, due either to concomitant mis-folding of GFP, or to aggregation of the mis-folded proteins, or to rapid degradation of the mis-folded proteins by the proteasome; similar mechanisms are likely to be at play with biotinylation of C-terminal BCCP tags.

Taken together, the advantageous features of the BCCP tag enable a facile route to microarray fabrication that circumvents the need for laborious pre-purification steps by combining in situ purification and immobilization into a single step: nanolitre volumes of crude insect cell lysates containing the expressed, BCCP-tagged proteins are printed directly onto streptavidin-coated hydrogel surfaces; the biotinylated, BCCP-tagged recombinant proteins bind with femtomolar affinity, essentially saturating the available streptavidin binding sites, irrespective of expression level; and all non-biotinylated proteins wash away, leaving a microarray of individually purified, BCCP-tagged human proteins tightly bound to the array surface [15]. Usefully, it has been observed by mass spectrometry that endogenously biotinylated host cell proteins (which can be readily observed on a western blot) do not pull down as
native proteins because their biotin moieties are not physically accessible to bind to streptavidin [113]. Furthermore, the streptavidin-biotin interaction is essentially irreversible ($K_d \approx 10^{-15}$ M) and the streptavidin tetramers are themselves covalently bound into the hydrogel surface, so no lateral diffusion of the arrayed proteins occurs, even after months of storage.

A further notable feature of the Immunome protein array platform is the use of more sophisticated surface chemistry, with polyethylene glycol (PEG) polymers serving two principle functions on the surface: highly effective blockade of non-specific macromolecule absorption to the underlying glass [22]; and provision of an environment surrounding the arrayed proteins that is similar to unstructured water and which helps to preserve the folded structure of the immobilized proteins for a long period (Sengenics claim a shelf life of 6 months for the immunome protein arrays). The underlying conformational flexibility of the PEG polymers also allows individual protein molecules in the same spot to orient themselves in 3D such that they can interact, enabling homo- and hetero-dimerization to occur (the latter, where different proteins have been co-arrayed). For example, the Blackburn group have shown that arrayed human protein kinases can autophosphorylate on the array surface in the presence of ATP, implying that catalytically-active homodimers have formed [15].

The Blackburn group has also developed a new bioinformatic pipeline for raw data processing and normalization of protein microarray data [2]. This pipeline includes flagging of saturated signals (≥10% saturated pixels, which represent signals above the reading capacity of the microarray scanner); background subtraction, which calculates net intensities by subtracting the neighborhood background intensities from matching foreground intensities; noise threshold, which discards net intensities that are less than two standard deviations from the mean neighborhood background intensity, that is, which are not significantly different from background therefore considered to be in the noise; percentage of coefficient of variation (CV) calculations, which determines the variation between replica features:

$$CV\% = \frac{\sigma}{\mu} \times 100,$$

where $\sigma$ is the standard deviation, and $\mu$ is the mean signal of all feature replicates; spot filtering, where spots with CV above 20% are flagged; mean of net intensity, which calculates the mean of all valid replicas per feature; verification of positive controls by CV evaluation, which confirms that array printing was successful and of good quality; and data normalization, which uses a composite normalization method that is based only on the positive control features (e.g. fluorescently-labeled and biotinylated BSA) [2].

Downstream of the raw data extraction and processing steps, for biomarker discovery assays, a penetrance fold change-based method is then used to identify up-regulated and down-regulated biomarkers when conducting a group analyses (e.g. a diseased cohort in comparison to an age- and gender-matched healthy cohort) and data is visualized using volcano plots, while receiver-operator characteristic (ROC) curves are also used to identify candidate diagnostic panels.

The Immunome protein array has been used for immune profiling across numerous disease settings, including systemic lupus erythematosus [103], malaria [104], Parkinson’s disease [105], melanoma [21,114], colorectal cancer [114], prostate cancer [32], non-small cell lung cancer [41] and ageing (submitted). Further applications of the Immunome protein array include cancer and autoimmune disease biomarker discovery, monitoring global immune responses to novel therapeutics (e.g. NY-ESO-1 vaccine [21], anti-CTLA-4, and/or anti-PD-1 immunotherapy [Duarte et al. submitted]; BRAK/MEK inhibitors [114]) in a clinical trial setting, and monitoring global immune responses to viral, microbial and fungal infections. In addition, kinase inhibitor selectivity assays have been reported across 336 human kinases on the Immunome protein array platform [15], with the microarray data matching previously reported solution-phase $K_i$ values well (Peton et al. submitted), supporting the contention that the proteins on this array platform are folded and functional.

Extensive linearity and dynamic range assays have been performed for this array, reporting a detection limit in the pg/mL range for autoantibody binding and a linearity of detection over ~4 orders [21]. Coupled with very low inter-array variability (inter- and intra-array CVs <4%), this means that the Immunome protein array platform is genuinely quantitative – making it the first full length protein microarray platform for which this is true. Equally importantly, compared to other platforms, the Immunome protein array gives sparse signals in serological assays on young, presumed healthy individuals [32,114], yet gives positive, specific binding data with conformation-specific antibodies. Together, this suggests that preservation of folded structure on the Immunome protein arrays – which is essentially a function of the underlying surface chemistry – should result in fewer false positives (since hydrophobic epitopes remain buried) and fewer false negatives (since discontinuous epitopes are preserved) than typically observed on other protein microarray platforms.

7. Conclusion

In this review, we have provided an overview of the technological challenges encountered by researchers in the protein microarray field regarding content generation, immobilization strategies, assay development, as well as data processing and analysis; we have then discussed technological advances to overcome some or all of these challenges in the context of the most commonly used protein microarray platforms (Table 1). To date, the greatest efforts have understandably been made in expanding the content on the various protein microarray platforms, with less attention being paid yet to the development of improved surface chemistries that will aid preservation of folded structure function of the arrayed proteins, or to the development of a diverse repertoire of quantitative, functional assays that are compatible with the various microarray platforms. Data handling is also evolving, with increasingly sophisticated data processing and statistical methods being used now by many groups, but standardization of appropriate statistics and of reporting remains an on-going issue.
At present, the dominant biological applications for protein microarrays are based on qualitative profiling of antibodies across repertoires of potential antigens. It seems intuitive that, when conducting antibody profiling, biomarker discovery or protein functionality studies, having full-length, correctly folded proteins on the array would be more desirable than shorter protein fragments or unfolded proteins. This goal presents a different magnitude of technological challenges, yet there are encouraging signs that this challenge is being addressed now, at least for some of the platforms described earlier, suggesting that functional assays on protein microarrays may become more mainstream in biology in the near future, if costs of access can be kept reasonable.

8. Expert commentary

Protein microarray technologies continue to develop, with the ultimate aim of full proteome coverage. To date the highest density protein microarrays have reached the point where at least one epitope per protein is present for 19,497 of the estimated ~20,500 human genes and it seems likely that every human protein will soon be represented by one or more epitopes on such arrays. Full length human protein arrays are not far behind, with current versions representing the expressed products of up to 15,000 unique protein coding genes. However, more generally the human protein microarray field will take longer to tackle questions of polymorphic variation, splice variation and dynamic post-translational modifications – issues that the mass spectrometry-based proteomics community are now turning their attention to.

Fundamentally though, irrespective of the available content, any protein microarray is only as good as the range of assays that can be performed on-array and it is this aspect that we believe still today provides one of the greatest barriers to broader uptake of this technology; this is evidenced by the relatively poor uptake of the yeast proteome microarray described herein, despite it containing ~80% of the theoretical proteome of the work-horse, model organism, *S. cerevisiae*. We suggest that there are several distinct ways to expand the menu of compatible assays, including *inter alia*:

(i) Broaden out the effective proteome coverage of existing on-array assays by introducing on-chip specific post-translational modifications of the existing content (e.g. phosphorylation; citrullination; methylation; acetylation; glycosylation; etc.) prior to functional assay, thereby enabling the regulation of protein activities to be studied in higher throughput;

(ii) Build protein complexes on the array surfaces prior to functional assay, either by expression of the content in homologous hosts (essentially equivalent to mass co-immunoprecipitation experimentiation to form the arrays) or through co-expression of known interacting partners in the context of the most commonly used protein microarray platforms (Table 1);

(iii) Broaden out the repertoire of on-array assays to include high content, cell-based readouts;

(iv) Incorporate neoantigens onto the arrays, not just wild type proteins, including aberrant splice variants and mutations; and

(v) Move away from reliance on fluorescence as the readout of functional assays on the arrays and into the realm of label-free detection of biomolecular interactions.

The first three of these possibilities will require a concomitant greater focus on underlying surface chemistries and on
preservation of the folded structure of proteins post-immobilization, since non-specific modification of arrayed proteins, non-specific complexes or non-specific cell binding to the surface would all obfuscate biologically meaningful, true positives in subsequent assays. Here, it is clear that simple microarray surfaces are not best and we suggest that the protein microarray field will do well to learn from other biosensor platforms, where site-specific attachment of capture probes to more complex, hydrogel-type surfaces is commonly used, albeit with much lower densities of content per chip, and where a broader range of biomolecular interactions are routinely probed on the surfaces.

The importance of neoantigens in immunology is now widely recognized in the cancer and vaccine fields, where accumulation of neoantigens appears to correlate with, for example, escape from immunotherapy or prophylaxis. It would therefore be fascinating to use protein microarray technology to explore the connection between tumor specific mutations and the development of corresponding mutation-specific autoantibody responses. This concept is not without its difficulties, since the recent literature on intra-tumor heterogeneity shows that the precise location of tumor biopsy sampling can have a significant effect on the spectrum of mutations identified by sequencing, which may in turn confound efforts to correlate protein array and sequencing data, but early papers in this area [115] suggest that there is definitely scope for progress. If such correlations can be established, then the prospects for use of focused protein microarrays to monitor treatment response and the emergence of resistance through serology seem bright. However, given that tumor-associated mutations can affect folding of the antigen and that autoimmune responses can occur to discontinuous as well as linear epitopes, we suggest that arrays of folded proteins are likely to be most useful here.

In terms of the development of new array-based assays, in principle any assay that can be conducted on a purified protein in solution ought to be feasible on a protein microarray, with some ingenuity. However, a major challenge is how to simultaneously visualize, for example, the binding of unlabeled small molecules to many different arrayed proteins, either from the same or unrelated families. Ideas from the activity-based protein profiling field [116,117] can be adapted to a degree, but such probes are typically limited to subsets of a given protein family and do not routinely give quantitative readouts. We therefore suggest that label-free detection methodologies, such as surface plasmon resonance, biolayer interferometry and mass spectrometry, hold greater promise for the future of protein microarray technology: by removing the requirement for a fluorophore to be present somewhere in the assay system for signal detection, quantification of biomolecular interactions across a broader range on structurally unrelated proteins on a protein microarray should then be enabled.

As an aside, it is interesting to reflect on the fact that, whereas a DNA microarray-based transcriptomic experiment typically returns a data point for every probe on the chip, a protein microarray experiment ought to only return data on a small subset (unless the assay is dominated by non-specific interactions). The current prices of commercial protein microarrays appear to be set based on some composite factor of the cost of content production and comparison to what researchers have paid for DNA microarrays in the past, yet the cost-per-resultant true datapoint is vastly different between DNA and protein microarrays. We therefore suggest that, despite the march toward complete proteomes on a chip, there will remain a demand for smaller, more focused protein microarrays that are customized to the specific research question and which can deliver relevant quantitative data for all the probes on the array.

Protein microarray data processing and analysis tools have come a long way, with current approaches starting to use more appropriate statistical methods. However, we believe that the restricted ability to compare datasets across experiments done in different laboratories and on different platforms represents another significant bottleneck in the field today. We therefore suggest that some degree of standardization of the statistical principles that are appropriate for protein microarray datasets, as well as of the reporting and annotation of data – as has happened in the DNA microarray, proteomics and glycan microarray communities through the MIAME [118,119], MIAPE [120], and MIRAGE [121] initiatives – would be beneficial to the protein microarray community now.

It is worth pausing briefly to consider the elephant in the room in any discussion on protein microarrays: the question of how the information that can be gained from protein microarrays compares to mass spectrometry-based proteomics. Mass spectrometry provides the ability to study given proteomes in both unbiased data dependent and data independent modes, as well as in targeted modes, and can detect PTMs, splice variants, and neoantigens. However, due to high sequence homologies, one class of proteins that mass spectrometry-based proteomics continues to struggle with is the immunoglobulins, so for serological studies that aim to quantify changes in the titres of individual (auto)antibodies in disease, the component-resolved protein microarray architecture has considerable advantages, both for discovery as well as for quantitative profiling applications.

The advent of chemical proteomic strategies that, for example, utilize crosslinking approaches to identify components of complexes [122] or immobilized chemical probes to detect potential drug binding proteins [123], have opened up new horizons for mass spectrometry-based proteomics. However, these mass spectrometry-based methods are not readily amenable to direct, high throughput quantitation of biomolecular interaction (e.g. protein–protein, protein–ligand, protein–DNA) affinities, which is crucial to any understanding of biomolecular selectivity, nor, for example, to the high throughput study of protein-cell interactions or the activities of protein complexes.

Thus, we suggest that both protein microarrays and mass spectrometry-based proteomics have their own niches, with protein microarrays having distinct advantages for systems-oriented, quantitative analysis of the functions of full length, correctly folded proteins and of protein complexes. We also suggest though that the convergence between these two proteomic technologies is likely to create a rich seam in the future, with mass spectrometry identifying new targets to put on protein microarrays for quantitative analysis and with protein microarrays becoming mass spectrometry-readable for analysis of mixtures.
This review has highlighted the fact that protein microarrays can be used across numerous applications. Of particular current interest globally is the use of protein microarray technology in biomarker discovery or immune response profiling in disease settings. The various protein microarray platforms described here have been used to identify sets of candidate diagnostic biomarkers in cancers, autoimmune diseases, inflammatory disease, neurological diseases, and chronic diseases. In the cancer biomarker discovery field, we believe that cancer-specific microarrays hold great promise in identifying novel diagnostic, prognostic, and disease recurrence biomarkers, as well as for monitoring immune responses to novel immuno- and chemotherapeutics in a clinical trial setting. However, as will the vast majority of proteomics data, the clinical validation and application of such candidate biomarkers is still a work in progress and, in our experience, a closer link between clinicians and scientists will be critical in order to begin translating interesting research observations made using protein microarray technology into clinical practice. In particular, once the discovery phase of research has been completed, identified candidate diagnostic biomarkers then need be correlated with detailed clinical knowledge about the individual patients and samples (including precise disease classifications at the time point at which samples were drawn) and then validated across large patient cohorts and available databases before being developed onto a suitable diagnostic platform for patient stratification, diagnosis and treatment monitoring.

9. Five-year view

In the next five years, we anticipate that several protein microarray platforms will achieve full coverage of the basic ‘one-gene-one-protein’ human proteome as well as complete proteomes for a number of other model organisms. We further anticipate that greater emphasis will be placed on broadening the proteome coverage of the protein microarrays to include post-translational modifications, splice variants, neoantigens and protein complexes. Greater focus will also be placed on surface chemistry to retain the folded structure of the arrayed proteins, in order to reduce false positives and false negatives in assays; these advances will collectively impact significantly on the utility of protein microarrays in downstream biomarker discovery and functional analysis applications. Smaller, more focused protein microarrays will be customized for specific studies and the cost of protein microarrays will be lower than today.

A key technological advance will be integration of protein microarrays with on-array cell-based readouts and with label-free detection methodologies, which will open out a much wider range of functional assays that can be performed and which will therefore significantly increase the relevance of protein microarray technologies to the study of complex biological phenomena in both humans and in model organisms such as yeast.

Translation into the clinic of (auto)antibody-based biomarkers discovered using protein microarrays will offer the prospect of pre-symptomatic diagnosis in cancers and of patient stratification in precision medicine.

Key issues

- Protein microarrays are approaching full coverage of the basic ‘one-gene-one-protein’ human proteome now and the complete proteome microarrays have also been produced for a number of smaller, model organisms, including yeast and _E. coli_.
- The major applications of protein microarrays currently lie in antibody screening and in serological profiling, where the component-resolved protein array-based methods have obvious advantages over mass spectrometry-based proteomics.
- Demonstrated protein microarray-compatible assays include on-array quantitation of protein-drug, protein-DNA and protein-protein interactions, as well as enzymatic activity assays, suggesting that protein microarrays are capable of generating high throughput, quantitative data on the selectivity and affinity of biomolecular interactions; this represents another area where array-based methods have obvious advantages over mass spectrometry-based proteomics.
- Formation of protein complexes and post-translational modifications on-array has also been demonstrated on some platforms, suggesting that protein microarrays have potential to aid the study of regulation in biological systems.
- In general, greater emphasis is needed on surface chemistries that are able to preserve folded structure and function of arrayed proteins, in order to underpin the development of a wider repertoire of high throughput, on-array functional assays.
- Integration of protein microarrays with cell-based readouts and label-free detectors is needed now to take the field beyond fluorescence-based assays.
- Smaller, focused custom protein microarrays continue to have a niche for specific biological questions, for example in drug selectivity screening or in vaccine design.
- Standardisation of bioinformatic methods, as well as for reporting and annotation of data for protein microarrays is needed now to enable datasets from different studies to be mined more effectively.

Funding

This paper was funded by a grant from the National Research Foundation of South Africa (64760).

Declaration of interest

JM Blackburn is a consultant for Sengenics and developed their Immunome protein array product while an academic researcher at the University of Cambridge. The Immunome protein array is one of the five commercially-available protein array platforms described in this review. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

ORCID

Jonathan M. Blackburn  [http://orcid.org/0000-0001-8988-9595](http://orcid.org/0000-0001-8988-9595)
References

Papers of special note have been highlighted as either of interest (●) or of considerable interest (★) to readers.


- Describes a generic pipeline for processing raw protein microarray datasets.


- Describes the first in situ generated protein array approach.


- Screening of a yeast proteome microarray for protein and phospholipid binding functions.


• Serological analysis of 60 pancreatic cancer, 51 ovarian cancer and age-matched controls on a ProteinArray identified 29 and 202 autoantibodies in the two cancers, respectively.


• Construction of an array of 16,368 unique full-length human ORFs.


• Identification of >ca. 18,000 protein–DNA interactions between 460 DNA motifs and 4191 human proteins of various functional classes, including a large number of unanticipated protein–DNA interactions.


• Protein–protein interactions, small-molecule–protein binding, protein kinase substrate identification and c-di-GMP binding proteins identified on a bacterial proteome microarray.


105. Induction of elevated autoantibody levels against 24 human proteins as a result of a Plasmodium infection.
107. Serological screening of 30 Helicobacter pylori-seropositive Parkinson’s disease patients and matched controls identified 13 significant autoantibodies, including against proteins essential for normal neurological functions.
116. Anti-p53 autoantibodies identified in pancreatic, ovarian, and breast cancer patient sera showed a polyclonal response to common epitopes shared between p53 point mutant proteins.
120. Quackenbush J. Data reporting standards: making the things we use better. Genome Med. 2009;1(11):111.1–111.3.