Chapter 3
Protein Function Microarrays: Design, Use and Bioinformatic Analysis in Cancer Biomarker Discovery and Quantitation

Jessica Duarte, Jean-Michel Serufuri, Nicola Mulder, and Jonathan Blackburn

Abstract Protein microarrays have many potential applications in the systematic, quantitative analysis of protein function, including in biomarker discovery applications. In this chapter, we review available methodologies relevant to this field and describe a simple approach to the design and fabrication of cancer-antigen arrays suitable for cancer biomarker discovery through serological analysis of cancer patients. We consider general issues that arise in antigen content generation, microarray fabrication and microarray-based assays and provide practical examples of experimental approaches that address these. We then focus on general issues that arise in raw data extraction, raw data preprocessing and analysis of the resultant preprocessed data to determine its biological significance, and we describe computational approaches to address these that enable quantitative assessment of serological protein microarray data. We exemplify this overall approach by reference to the creation of a multiplexed cancer-antigen microarray that contains 100 unique, purified, immobilised antigens in a spatially defined array, and we describe specific methods for serological assay and data analysis on such microarrays, including test cases with data originated from a malignant melanoma cohort.

Keywords Protein microarrays • Cancer–testis antigens • Cancer biomarker discovery • Bioinformatic analysis • Pipeline

J. Duarte • J.-M. Serufuri • N. Mulder • J. Blackburn, D.Phil (✉)
Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, N3.06 Wernher Beit North Building, Anzio Road, Observatory, Cape Town 7925, South Africa
e-mail: jonathan.blackburn@uct.ac.za

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_3,

3.1 Introduction

In the postgenomic era, attention has turned towards the systematic assignment of function to proteins encoded by genomes. Bioinformatic methods are typically now used ubiquitously as an essential first step in assigning predicted function to open reading frames (Hunter et al. 2009). However, while such methods give helpful insights into possible function, there remain many examples of proteins that have closely related sequences and/or structures but which prove to have quite different functions when studied experimentally (Wise et al. 2002; Schmidt et al. 2003; Lander et al. 2001). As the number of sequenced genomes expands ever further, there is thus an ever-increasing need for experimental methods that enable the determination and/or verification of protein function in high throughput. At the forefront of this monumental task, the field of proteomics can be segregated into discovery- and systems-oriented proteomics (Macbeath 2002). Discovery-oriented proteomics is mainly concerned with documenting the abundance and localisation of individual proteins as well as building a picture of protein–protein interaction networks. This is the realm of 2-hybrid screens, 2D-gel electrophoresis and increasingly powerful more direct, isotope-labelling-based mass spectrometry methods; these latter two methods in particular are commonly used to understand the way in which expression profiles change in response to different stimuli by comparing, for example, diseased and healthy cell extracts. However, these discovery-oriented proteomic methods tell us little directly about the precise function of individual proteins or protein complexes, even when augmented by ever more sophisticated bioinformatic methods. Systems-oriented proteomics takes a different approach; rather than rediscovering each protein in each new experiment, the focus is on a predefined set of proteins – in principle up to an entire proteome, but in practice more typically a limited subset thereof – enabling the functionality of each member of that set to be dissected in great detail (Wolf-Yadlin et al. 2009). However, obtaining quantitative and genuinely comparative functional data across large sets of proteins with any degree of accuracy is technically difficult, requiring isolation of each individual protein in an assayable format. We and others have chosen to focus on protein function microarray-based methods because the parallel, high-throughput nature of microarray experiments is attractive for analysing large numbers of protein interactions, while the uniform intra-array conditions both simplify and increase the accuracy of assays (Wolf-Yadlin et al. 2009; Boutell et al. 2004; Kodadek 2001; Predki 2004; Zhu et al. 2000, 2001; Michaud et al. 2003; Fang et al. 2003). Additionally, the small volumes of ligand or reaction solution required to perform assays, typically tens to hundreds of microlitres, can provide economic advantages, for example, when using expensive recombinant proteins or labelled compounds.

The key element to such microarray experiments is that the arrayed, immobilised proteins retain their folded structure such that meaningful functional interrogation can then be carried out. There are a number of approaches to this problem, which

differ fundamentally according to whether the proteins are immobilised through non-specific, poorly defined interactions or through a specific set of known interactions. The former approach is attractive in its simplicity and is compatible with purified proteins derived from native or recombinant sources (MacBeath and Schreiber 2000; Angenendt et al. 2003) but suffers from a number of risks. Most notable among these is that the uncontrolled nature of the interactions between each protein and the surface might at best give rise to a heterogeneous population of proteins or at worst destroy activity altogether due to partial or complete surface-mediated unfolding of the immobilised protein. In practice, an intermediate situation probably most often occurs, where a fraction of the immobilised proteins either have undergone conformational change as a result of the non-specific interactions or have their binding/active sites occluded by surface attachment; these effects effectively reduce the specific activity of the immobilised protein and therefore decrease the signal-to-noise ratio in any subsequent functional assay that is sensitive to conformation. It is, therefore, important to consider the possible effects of unfolding on the intended downstream assay prior to choosing an array surface: for example, an assay in which a solution-phase kinase phosphorylates arrayed proteins may well be sensitive to disruption of the relative three-dimensional arrangement of targeting and substrate domains in the arrayed proteins (Blackburn and Shoko 2011); by comparison, an assay in which solution-phase antibodies bind to linear epitopes on the array will be unlikely to be affected by unfolding of the arrayed proteins – indeed, it may even be desirable to deliberately unfold such proteins in order to expose a greater range of potential epitopes. However, an important caveat here is that unfolded proteins are also more likely to exhibit non-specific binding to antibodies via the now-exposed hydrophobic surfaces and, across an array of diverse proteins, this non-specific binding may give rise to a high rate of false positives in serological assays.

The advantages of controlling the precise mode of surface attachment are that, providing the chosen point of attachment does not directly interfere with activity, the immobilised proteins will have a homogeneous orientation resulting in a higher specific activity and higher signal-to-noise ratio in assays, with less interference from non-specific interactions (Koopmann and Blackburn 2003). This may be of particular advantage when studying protein–small-molecule interactions or conformationally sensitive protein–protein interactions in an array format. The disadvantages of this approach though are that it is really only compatible with recombinant proteins or with families of proteins, such as antibodies, which have a common structural element through which they can be immobilised. However, in a systems-oriented approach, the disadvantage of working with recombinant proteins is largely outweighed by the problems encountered in individually purifying large numbers of active proteins from native sources. In addition, experimental approaches that facilitate high-throughput expression and purification of many different proteins in parallel have become more generally accessible over recent years, simplifying access to larger, defined collections of recombinant proteins. An important caveat here though is that it is increasingly clear that despite its ease of use, *Escherichia coli* is not an optimal host for

recombinant expression of folded, functional mammalian proteins. Furthermore, while cell-free transcription/translation-based protein microarray systems have been described (He and Taussig 2001; Ramachandran et al. 2004), it remains unclear how reproducible such arrays are or what proportion of mammalian proteins produced by such approaches are properly folded and therefore functional prior to immobilisation (Blackburn and Shoko 2011) – thus, their true utility in cancer biomarker discovery applications also remains unclear.

Having decided on and created the protein content and then fabricated a reproducible protein microarray, a typical downstream protein microarray experiment generates a large amount of raw and often noisy data that requires preprocessing via a series of computational steps in order to filter and normalise the data to yield a robust clean data set which can then be analysed using various bioinformatic tools in order to determine its biological significance (Klein and Thongboonkerd 2004; Brusic et al. 2007). However, DNA microarray software solutions in general have proved to be not well suited to the analysis of protein microarray data – for reasons that will be discussed below – and a comprehensive analysis software solution designed for protein microarrays has yet to be produced. To address this issue, our group has developed a preprocessing and quality control pipeline for raw protein microarray data (Zhu et al. 2006) which we will discuss in more detail below.

3.2 Custom Antigen Arrays as Serological Diagnosis Tools

A number of different types of protein microarrays are currently used to study the biochemical activities of proteins. Among them, the analytical microarrays are typically used to profile complex mixtures of proteins and to estimate the level of expression, binding affinities and specificities of specific components. These types of protein microarray-based experiments have interesting applications in molecular medicine, such as biomarker discovery for diagnosis, drug design and development as well as increasing understanding of pathogenesis and disease biology (Hall et al. 2007).

In the cancer field, it is well understood that individual patients generally show aberrant expression and/or post-translational modification of a selection of antigens (‘cancer antigens’), suggesting that detection and quantitation of cancer antigens should be a promising approach in, for example, disease diagnosis. In an ideal world, cancer biomarkers should be easy and inexpensive to measure to allow easy accessibility in developing countries, should be measureable in peripheral fluids (e.g. serum) to avoid unnecessarily invasive treatments and should be highly accurate for diagnostic and/or prognostic purposes to allow improved clinical management of patients (Berrade et al. 2011; Frank and Hargreaves 2003; Rifai et al. 2006).

Considering serum as a target peripheral fluid for cancer diagnosis, the first step today typically involves identification through serology (using the term in its broader

sense) of a set of candidate serum biomarkers that might correlate with disease status. In any serological analysis of cancer patient samples, there are two complementary approaches that could be taken: one obvious approach is to identify and quantify circulating tumour antigens that are aberrantly expressed in cancer; however, it is also known that aberrant protein expression and/or post-translational modification results in measureable autoimmune responses in most cancers, thus providing a second approach to serological analysis based on the identification and quantitation of circulating autoantibodies to tumour antigens.

Circulating tumour-antigen profiles may be analysed either by direct mass spectrometry-based proteomic techniques or by using antibody microarrays made up of immobilised antitumour-antigen antibodies. Alternatively, circulating autoantibody profiles can be analysed using protein microarrays made up of immobilised tumour antigens. There are pros and cons to each approach. Mass spectrometry approaches to de novo discovery of serological markers have not met with much success to date, largely due to the complexity and dynamic range of the serum proteome. Selected reaction monitoring (SRM) mass spectrometry-based assays that target known antigens are likely to be more successful but still today require significant preprocessing of serum samples (e.g. depletion of abundant proteins, tryptic digests and LC separations) that may influence downstream quantitative assays.

Antibody microarrays provide a much more direct means to analyse a wide range of different cancer antigens in what amount to miniaturised, multiplexed ELISAs. Numerous antibody microarrays are now available commercially and are being developed to be able to both quantify the individual targeted antigens and to also assess changes in post-translational modifications (e.g. glycosylation) of a given antigen between samples. The obvious current limitation in antibody microarray technology is the availability of large collections of suitable high specificity anticancer-antigen antibodies; this shortcoming is being addressed via a number of public initiatives (e.g. the Human Protein Atlas (www.proteinatlas.org), which currently describes antibodies to 12,238 human proteins). However, a less obvious limitation is that the antibodies used in antibody microarrays are typically murine in origin and have binding affinities for their target antigens in the nanomolar range, which limits the ability of antibody microarrays to detect circulating tumour antigens at the sub-nanomolar concentrations likely to be present at low tumour loads, that is, in early disease; this is less easy to address and may well prove to be a limitation for antibody microarrays into the future (note that in any such antibody-based microarray assay, the array-based signal is dependent both on the analyte concentration and also on the antibody–analyte affinity due to simple equilibrium binding considerations. Thus, as the analyte concentration significantly falls below the K_d of the antibody–analyte binding interaction, the proportion of immobilised antibody that is bound by analyte – and therefore the signal from the array – becomes nonlinear and falls off rapidly into the background noise).

Antigen microarrays by comparison seek to detect and quantify circulating cancer-specific human autoantibodies, which are typically highly specific for the tumour antigen and show picomolar or lower affinities. Antigen microarray-based serological analyses thus offer unique opportunities to find novel disease-specific serological

markers – that is, cancer-specific autoantibodies – that are not readily accessible by other proteomic technologies (Matarraz et al. 2011). More specifically, the study of autoantibody profiles in cancer patients could lead to the discovery of novel biomarkers for, inter alia, early detection of tumours, patient stratification, personalised patient treatment, development of improved therapies, and monitoring therapeutic response and disease progression. Importantly, antigen microarrays provide the enticing possibility of detecting much lower concentrations of autoantibodies than are detectable for the cancer antigens themselves (due to the intrinsically high affinity of human autoantibodies), thus in principle bringing forward in time the ability to detect the cancer biomarkers and potentially enabling presymptomatic diagnosis. As with antibody microarrays, a limited number of antigen arrays are now available commercially (e.g. the Invitrogen Human ProtoArray which contains ca. 9,000 individually purified, denatured human proteins).

However, antigen microarrays also suffer problems, primarily due the fact that often the aberrant form of any given cancer antigen that is misrecognised by the host immune system as an autoantigen is poorly defined, making it difficult to create recombinant forms of the autoantigen for reproducible protein array fabrication. The choice and customisation of the antigen content on such microarrays is thus a critical component in experimental design and will strongly influence the likelihood of downstream success (Zhu et al. 2006; Ingvarsson et al. 2007; Hultschig et al. 2006; Sanchez-Carbayo 2006; Casiano et al. 2006). One group of cancer antigens – the so-called cancer–testis antigens – therefore stands out as being of particular potential utility in serological analyses of cancers.

The cancer–testis (CT) antigen family are a group of >90 structurally and functionally unrelated proteins that show highly restricted expression only in germ cells in the adult male testis, as well as in occasionally in the adult ovary and the trophoblast of the placenta (Scanlan et al. 2002). Critically, these are all immune-privileged compartments, so the adult immune system has typically never been trained to recognise CT antigens and ‘self-antigens’. The CT antigens are however aberrantly expressed in many cancers due to the disruption of gene regulation. When this occurs these proteins are thus misrecognised as autoantigens, making them potential cancer diagnostic markers as well as vaccine targets.

The expression of different CT antigens is known to be associated with many different types of cancer. However, the absence or presence of expression of any one CT antigen is not in itself exclusively indicative. Thus, as with many other candidate biomarkers, CT antigens are therefore not thought to be individually viable as diagnostic or prognostic markers (Scanlan et al. 2002; Anderson and LaBaer 2005). Instead, it is possible that patterns of CT antigen expression may be used as correlates of specific cancer types and of disease progression, making this family of cancer antigens particularly well suited to serological study via an antigen microarray-based approach. Importantly, recombinant forms of CT antigens typically retain immunological activity, further enhancing their suitability as content for customised cancer-antigen microarrays.

In contrast to systemic autoimmune diseases, where the presence of a single autoantibody might have a diagnostic value, tumour-associated antibodies have little diagnostic value when detected individually for three reasons: (1) the frequency of a specific antigen within a cohort of patients is often relatively low; (2) several tumour-associated antigens are responsible for tumorigenesis in multiple cancer types, so the detection of the associated autoantibody can only indicate the presence of a developing tumour without enabling discrimination between different cancer types; and (3) several CT antigen-associated autoantibodies lack specificity, as they might arise from events associated with other diseases. We therefore concluded that the characterisation of autoantibody profiles against a wide panel of CT antigens via a CT antigen array would be significantly more informative than the detection of autoantibodies against individual-specific antigens (Casiano et al. 2006; Robinson 2006). Our group has therefore developed a CT antigen array ('CT100 array') for the in vitro serological assessment of autoimmune responses to cancer-specific targets.

In seeking to exemplify the utility of our CT antigen array, we have focussed initially on the observation that an increasing number of clinical trials in progress today are examining the safety and efficacy of therapeutic cancer treatments which either target specific tumour-associated antigens (e.g. the CT antigens NY-ESO-1 and MAGE A3) or aim to transiently deregulate self-recognition of molecules by targeting components of the peripheral tolerance machinery such as CTLA4 (Ueda et al. 2003). However, in therapeutic vaccine trials, the chronic nature of the disease makes it difficult to provide an early assessment of whether an individual patient is generating a therapeutically useful response following vaccination or not. T-cell-based assays for cytokine release have been widely used in the search for correlates of therapeutic response, but so far without great success.

We have therefore applied our CT antigen microarray-based approach together with robust bioinformatic algorithms for data analysis, as a generic tool with which to study the serum antibody (i.e. B-cell) responses to therapeutic cancer experimental treatments, accepting implicitly that such analyses will only be a surrogate for the T-cell responses currently desired in cancer vaccine strategies. We will exemplify this approach below by reference to use of our CT100 microarray in the assay of serum samples from malignant melanoma patients who were undergoing an experimental vaccine treatment, our goal being to identify autoantibody 'biosignatures' that could serve as potential biomarkers of therapeutic response.

In this chapter, we provide an overview of the complete process necessary for creation and use of customised antigen microarrays in serological analyses. We consider general issues that arise in antigen content generation, microarray fabrication and microarray-based assays and provide practical examples of experimental approaches that address these. We illustrate these approaches by reference to serological assay of samples from a malignant melanoma cohort using a multiplexed cancer-antigen microarray that contains 100 unique, purified, immobilised antigens in a spatially defined array. We then focus on general issues that arise in raw

data extraction, raw data preprocessing and analysis of the resultant preprocessed data to determine its biological significance, and we describe computational approaches to address these.

3.3 Protein Microarray Technology

Fundamentally, protein microarray technology is based on the immobilisation of multiple proteins onto a surface (typically glass, gold or plastic) in a spatially defined array for use as capture probes. This technology permits numerous biological questions to be addressed via the high-throughput analysis of biochemical and/or biological interactions between the arrayed proteins and other biomolecules contained in complex sample solutions, thereby generating significant volumes of proteomic information that require analysis (Hardiman 2003).

The high-throughput manipulation of proteins to create protein microarrays is considerably more challenging however than the manipulation of oligonucleotides to create DNA microarrays; this is primarily due to the very diverse physical and chemical properties of different proteins, including differing stabilities, binding affinities and the requirement often of folded 3D structure for biological activity (Hardiman 2003; Draghici 2003). Considerations regarding protein acquisition, immobilisation and assay are discussed further below.

3.3.1 Antigen Content Generation

The methods used in antigen content generation usually rely on the identification of open reading frames (ORFs) and the selection of the most appropriate plasmids for protein expression. An ORF is a protein-coding segment within the DNA sequence that encodes all the amino acids between the initiation and termination codons. This ORF is typically amplified by PCR prior to insertion into an expression plasmid. In the case of eukaryotic genes that are subject to splicing, the protein encoding ORF is usually derived from cDNA – itself prepared from mRNA – to ensure that the correct protein sequence can be expressed in recombinant form in a suitable host (Hall et al. 2007; Phizicky et al. 2003; Gray et al. 1982). Heterologous expression is generally used, even though it may lead to expression issues. For example, in more than 60% of cases, soluble proteins expressed in *Escherichia coli* (*E. coli*) show altered solubility, suggesting folding defects, and lack any post-translational modifications. Insect cells by contrast provide a relatively simple eukaryotic alternative to *E. coli* and, importantly, utilise eukaryotic co-translational folding and post-translational modification pathways to typically yield properly folded forms of recombinant eukaryotic proteins with similar post-translational modifications to mammalian cells (Phizicky et al. 2003). In the protocols developed by our group, we

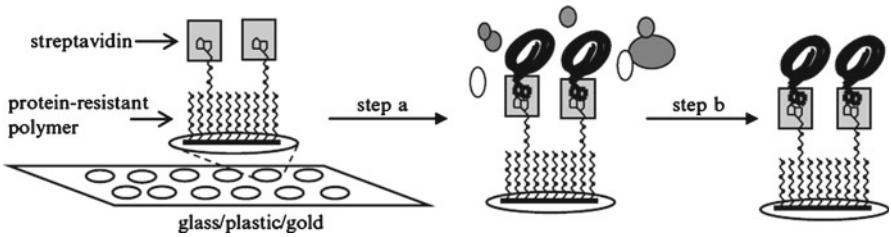

Fig. 3.1 Schematic of single-step immobilisation/purification route to array fabrication. The array surface is intrinsically 'nonstick' with respect to proteinaceous material but has a high affinity and specificity for biotinylated proteins. Crude cellular lysates containing the recombinant biotinylated proteins can then be printed onto the surface in a defined array pattern (*step a*) and all non-biotinylated proteins removed by washing (*step b*), leaving the recombinant proteins purified and specifically immobilised via the affinity tag in a single step

thus primarily use insect cells to express our human antigens of interest (Beeton-Kempen et al. submitted) (see Supplementary Material for detailed protocol).

3.3.2 Protein Immobilisation on Microarray Surfaces

The techniques of immobilisation are important both for effective concentration and orientation of immobilised proteins on the surface and also to preserve their folded conformations. There are two categories of protein immobilisation methods, covalent and non-covalent. The covalent immobilisation method is based on a covalent coupling to a cross-linker attached to the surface. By contrast, the biotin–streptavidin methodology used by our group is a non-covalent immobilisation method based on the high affinity of the biotin and streptavidin interaction. This method links biotinylated macromolecules to a surface that was previously derivatised with streptavidin via a single point of attachment (Fig. 3.1) (MacBeath and Schreiber 2000; Büssow et al. 2001) (see Supplementary Material for detailed protocol: Methodology, Sect. 1.4).

3.3.3 Detection of Binding Interactions on Microarrays

One of the main goals in protein microarray experiments is the quantitation of interactions between the probes immobilised on the slide surface and target analytes contained in the sample solution. To permit detection of this typically bimolecular interaction, molecules with specific interaction properties are labelled with either fluorescent, photochemical or radioisotope tags. The choice of one detection method over another depends on the need to reach a low signal-to-noise ratio at an affordable cost for the specific assay in question.

Since the vast majority of target analytes are not naturally coloured, fluorescent, bioluminescent or radioactive, detection of the molecular interaction on a protein microarray usually requires that some detectable molecule (a 'label') be included in the assay, either by direct conjugation to the analyte molecule itself or by conjugation to some secondary detection agent (e.g. an antibody) that can also bind to the analyte once it is specifically captured onto the array surface. Label-free biosensors (e.g. surface plasmon resonance- and quartz crystal microbalance-based biosensors) circumvent this problem, but in most manifestations are not yet truly compatible with medium- to high-density protein microarrays, plus they struggle to match label-based methods in terms of assay sensitivity (i.e. limit of detection), so will not be considered further here.

The choice of potential labels is wide: Chemiluminescence is a highly sensitive label-based detection method, but it has a relatively limited dynamic range; radioactivity-based methods are much less frequently used today because of safety concerns; fluorescent labelling is still therefore currently the most widely used detection method for protein microarray experiments since it is highly sensitive, stable, safe and effective and can be archived for future imaging. However, the direct labelling of analyte molecules might affect their ability to interact with their respective binding partners (Klein and Thongboonkerd 2004; Hall et al. 2007; Schweitzer et al. 2003). In the context of antigen microarray-based assays to measure human autoantibody profiles, the simplest mode of detection is thus via use of a fluorescently labelled antihuman IgG as a secondary detecting agent, since this will bind with high affinity to all human autoantibodies captured onto the antigen array surface, removing the need to label each target analyte in every biological sample.

3.4 Protein Microarray Data: Preprocessing

Even though microarray technologies have been around for many years, they are still subject to bias and variations. In addition to the variations introduced by probe acquisition, immobilisation and detection method, there are still a large number of environmental factors which might introduce variability in these experiments, including ambient conditions when the arrays were processed, the individual conducting the experiments, the recombinant sample differences, the variations in sample preparation, the nonuniformity in the hybridisation across an array surface, the distribution of artefacts or smears on the surface of the array and modifications in the scanner settings used for acquisition of fluorescent data. Nevertheless, a good experimental design can reduce noise and be beneficial for the downstream data analysis (Ingvarsson et al. 2007; Hultschig et al. 2006; Smyth and Speed 2003; Steinhoff and Vingron 2006; Altman 2005).

In addition to designing a standard optimised protocol to reduce noise within the microarray data, the experimental design should include the capacity to assess the quality of the microarray data, to filter out poor quality arrays and to normalise good quality data. With this consideration in mind, controls should be immobilised

onto the array surface, such as housekeeping and exogenous controls. Housekeeping controls are assumed to maintain a constant signal, either individually or collectively, within the different conditions of the experiments, while exogenous controls are those from species other than the one under study, which are generally selected to not give rise to a signal. Before proceeding with an intended array layout and assay, it is essential to identify a stable source of controls (Causton et al. 2004). In the context of an antigen microarray-based autoantibody profiling assay, simplistically one could consider arraying human IgG and sheep IgG as positive and negative controls.

3.4.1 Image Processing: Raw Data Extraction

Following a typical serological assay on an antigen microarray, in spite of several washing steps involved in our protocol, the spot quantification still remains affected by intangible factors. As illustrated in Fig. 3.2, the same lab protocol can lead to different results regarding the background of different replica arrays. The measures of background intensity usually represent the autofluorescence of the array surface at different spot locations (Causton et al. 2004).

The aim of image processing is to estimate the amount of specific anti-CT antigen autoantibody present in the serum by measuring pixel intensities across all probed spots. The image analysis software (ArrayPro Analyzer software (Media Cybernetics Inc., USA)) associated with the scanner (Tecan LS Reloaded fluorescence microarray scanner, Tecan Group Ltd., Switzerland) allows us to retrieve some statistics for the pixels measured in both the spots and their local backgrounds, for instance, the mean and median pixel distributions of the spots and their adjacent backgrounds. The scanner PMT (photomultiplier tube) gain setting helps discriminate between a weak signal and its background. By increasing the PMT gain setting, the sensitivity of the signal is improved but the selection of the optimal gain setting must provide a balance between the need to detect as many spots as possible and the necessity to avoid saturation of any of the spots.

The scanner software proceeds by matching a grid layout defined by the user (typically based on a .gal file generated by the microarray printer) to the actual image coming from the array and by locating the signal spots in order to quantify them. The spot finding can be achieved manually, automatically or semiautomatically. In the manual approach, the user adjusts a grid over the array and fits each spot individually to account for spot size variations and uneven spacing between spots. In practice, this approach is time consuming and subject to human error, especially when dealing with a large number of arrays. The automatic approach aims to identify and fit, without human intervention, the extent of each spot using specific algorithms. This approach is rapid and avoids human errors, but noise and contamination can lead to false detection of certain spots. The semiautomatic spot finding approach used by our group relies on automatic spot finding approach followed by manual curation of the grid alignment.

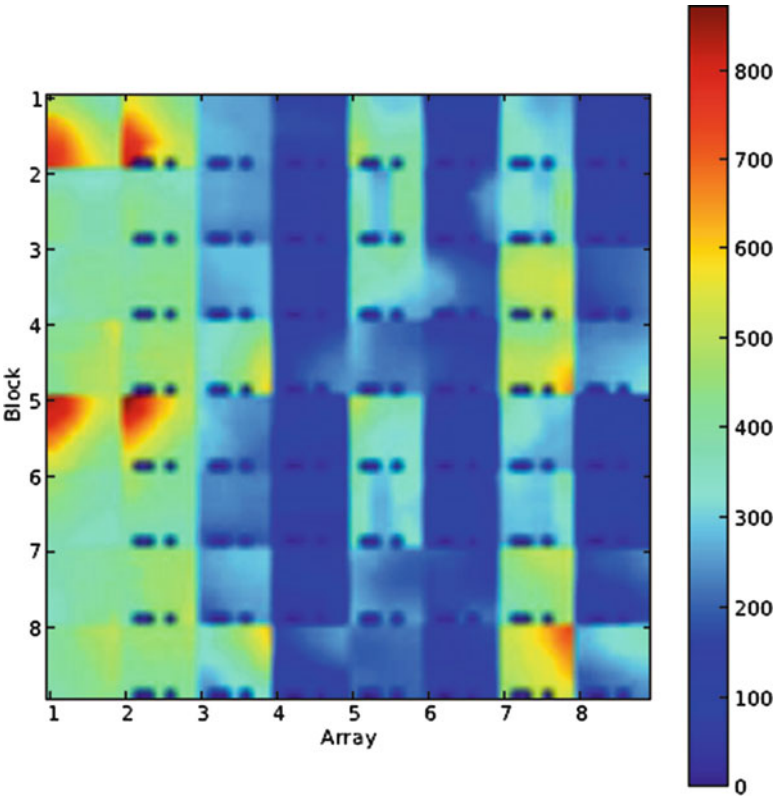

Fig. 3.2 Illustration of the variability of the background intensities of eight arrays with the application of the same experiment protocol. Here the eight blocks of each array are represented in the columns

3.4.2 Image Processing: Pixel Segmentation

Within a given spot area, individual pixels may be either true signal, background signal or erroneous signal originated from dust particles/artefacts. Typical image processing software outputs from scanned microarray images include some statistics based on pixel distributions. Image segmentation then aims to define rules to filter out unwanted pixels (see Fig. 3.3).

The simplest image segmentation rules include 'pixel filtering' and 'trimmed pixel' methods. The pixel filtering method sets an arbitrary threshold to filter out low-intensity pixels and performs statistics on the remaining pixels. The trimmed pixel method assumes that most of the pixels in the spot area are true signal, while most of those in the adjacent area belong to the background; for each spot or background intensities, the pixels falling outside defined quantiles are considered to be outliers and are therefore trimmed off (i.e. removed from further analysis) with all

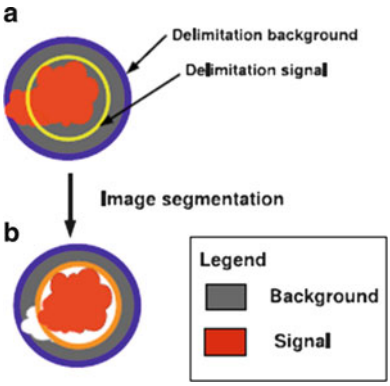

Fig. 3.3 Illustration of image segmentation. (**a**) shows the result of the spot finding process and (**b**) shows the result of image segmentation

subsequent statistics (e.g. mean and median foreground and background pixel intensities for each spot) being based on the remaining pixels. Significant differences between the effectiveness of the various segmentation methods become more noticeable as the level of artefacts on the arrays increases.

3.4.3 Image Processing: Quality Control

When evaluating the quality of an array image, there are many considerations that should be taken into account, such as:

1. Spot-to-spot variation – when looking at the signal of triplicate spots, one should expect a similar signal across all replicas, as well as uniform spots across the whole slide. Variations which may occur include spot bleeding (2 or more spots run into each other due to close proximity between spots or inappropriate spotting buffer, compromising the signal of all affected spots); pin sticking or erroneous pin calibration (inadequate cleaning of the arrayer pins and printhead between print runs could lead to the failure to print some or all intended spots, as well as non-reproducible printed spots); and washing artefacts and speckles (inadequate washing steps throughout the assay could lead to large washing artefacts which appear as negative spots or random additional smaller spots across the slide) on the array surface.
2. Spot homogeneity – when looking at the signal of an individual spot, one should expect a homogenous signal across all pixels within that spot. Variations which may occur include the 'doughnut' effect (inadequate pin height during print run and liquid residues on the pin body when immersing pin head in source plate could lead to uneven spot distribution); dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles on spots of interest, which appear as high-intensity pixels and skews the real signal) on the array surface; and temperature and humidity conditions (increased

temperature/decreased humidity may lead to the evaporation of printed spots, and humidity above 75% may lead to condensation, which would account for an uneven intensity within spots). The homogeneity between replica spots can be measured by calculating the coefficient of variation (CV), which is the ratio between the standard deviation (SD) of all pixel intensities within a spot and the mean intensity as a percentage. The CV should not exceed a value of 20%.

3. Background variation – when looking at several spots across an array, one should expect low variation of the background between the neighbourhood spots. Variations which may occur include dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles around spots of interest, which results in high local background for a specific spot and difficulty distinguishing between real signal) on the array surface.
4. Signal-to-noise ratio – when looking at several spots across an array, one should expect the spot intensity to be greatly above its local or neighbourhood background. Variations which may occur include washing artefacts and speckles (inadequate washing steps throughout the assay could lead to large washing artefacts which appear as negative spots or random additional smaller spots across the slide) and dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles around spots of interest, which results in high local background for a specific spot and difficulty distinguishing between real signal) across the array surface. To be confident that the net spot intensity (i.e. foreground intensity minus background intensity) is significantly above background, a signal-to-noise ratio of at least 2 is used for quality assurance, with 'noise' defined as the standard deviation of the background pixels.
5. Saturated pixels – when looking at several spots across an array, no saturated pixels should be visible within the spots of interest, as these are above the scanner's reading capacity (approximately 65,000 RFU in the case of the Tecan Reloaded scanner used in our laboratory). If this occurs, the slide should be rescanned at a lower PMT gain setting until no saturation is visible across the array (Schäferling and Nagl 2006; Espina et al. 2003).

3.4.3.1 Background Correction and Subtraction

Following pixel segmentation, the background intensity for a given spot is estimated from the adjacent area of that spot and is subtracted from its foreground intensity to obtain the net intensity, that is, the true signal derived from the specific interaction being interrogated on the array – in our case, the antigen–autoantibody binding interaction. The most important factor here is to avoid including artefacts in the estimation of the background because that could arbitrarily induce overestimated background intensity and, as a result, artificially reduce the true value of the spot intensity.

Module 1 of the protein chip analysis tools (ProCAT) (Zhu et al. 2006) provides a robust way to tackle the issue of local artefacts in background signals. The ProCAT approach for background correction essentially replaces the local

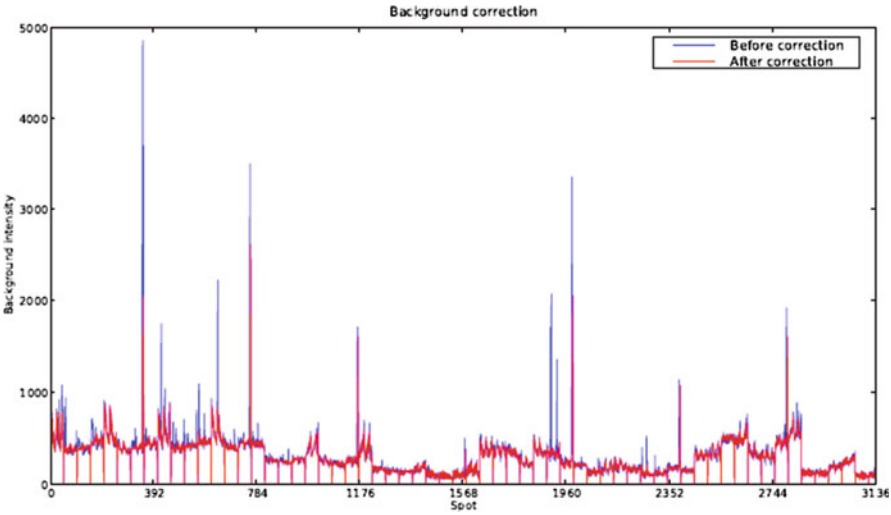

Fig. 3.4 The background correction corrects the arbitrary peak of the background intensity in a 3×3 spot window in a sample test case

median background intensity of a specific spot by the 'neighbourhood background', $0.5em\widehat{B}_{i,j}$, defined as the median background pixel intensity of a surrounding 3×3 spot window centred on the spot of interest:

$$\widehat{B}_{i,j} = \text{median}\{B_{i',j'}\}$$

$$i-1 \le i' \le i+1; j-1 \le j' \le j+1$$

where *i, j, i′* and *j′* design the row and column coordinates of spots.

This neighbourhood background correction smooths the local background (see Fig. 3.4) by reducing the effect of artefacts and noise in the background, thus enabling calculation of more accurate net intensities by subtracting the corrected neighbourhood background value from the median local foreground pixel intensity.

3.5 Protein Microarray Data: Filtering

Following background subtraction for each spot on the array, it is useful to then run a number of quality control (QC) tests to filter out noisy or defective array data prior to bioinformatic analysis. Data filtering therefore increases the data quality by flagging

questionable and low-quality arrays and/or individual spots. Our approach to this problem utilises a collection of criteria to filter out poor quality data. Among these are:

1. Flagging of spots with foreground intensity close to the saturation level (65,536RFU).
2. Flagging of triplicates based on their variability, as measured by the CV of their net intensities. N.B. When one of the triplicates is flagged, the measure of variability is then defined by the two remaining spots $S1$ and $S2$ as being equal to $(|S1 - S2|)/(S1 + S2)$.
3. Flagging of net signal intensities close to the noise level. A way to estimate the level of noise in the neighbourhood of a spot is to measure the standard deviation of the local background and to stipulate a noise threshold of 2SD of the local background pixel intensity (Tecan LS TM).

The negative controls on the array surface can in principle also enable the filtering of low-intensity spots because they reflect the cross reactivity of the detecting antibody with copurifying insect cell proteins or with the BCCP tag. However, in reality this typically proves less straightforward since both the expression level and the physical accessibility of the BCCP tag may differ in a difficult-to-quantify manner from antigen to antigen. It may therefore prove more effective in practice to define a baseline grass intensity signal for each antigen observed across a significant number of samples from healthy volunteers, when available.

3.6 Protein Microarray Data: Normalisation

Depending on the printing method used to fabricate the protein microarrays, each spot may or may not be printed with the same pin or nozzle. Where different spots are printed by different pins/nozzles, there may be subtle differences in the volumes of the printed antigens and therefore in the density of the immobilised antigens. Such differences can be corrected by so-called 'pin-to-pin' normalisation. More importantly, to correct for differences in the density of the same arrayed antigen across replica arrays, as well as to correct for any other systematic, non-biological variation between arrays, so-called 'array-to-array' normalisation is typically then applied.

The overall purpose of normalisation is thus to correct microarray data from variations in their measurements due to processes other than the targeted biological activity, thereby improving the quality of the data and allowing comparison of data originating from different arrays and different experiments (Lu et al. 2005; Oshlack et al. 2007).

3.6.1 DNA Microarray Normalisation Methods

Numerous methods for microarray normalisation have been published for 'two-colour' DNA microarrays. In two-colour arrays, two samples – control and target – are assayed on the same array with control and target analytes labelled with different fluorophores, simplifying downstream normalisation of the data. However,

in most cases protein arrays are run as single-colour assays due to concerns about differential physical occlusion effects arising from the size of the assessed macromolecules, together with the fact that the analyte molecules themselves are typically not directly labelled. Protein microarray-based data therefore typically lacks strong biological assumptions to carry out normalisation (Zhu et al. 2006). However, the methods published for normalising DNA microarrays can still inspire and support the normalisation of protein arrays and can be classified into three categories (Smyth and Speed 2003; Steinhoff and Vingron 2006; Freudenberg 2005):

1. *Scaling methods* assume that the arrays being normalised share a common statistical measure, such as the mean or median of their spot intensities or even the total intensities of their spots, and apply a common factor to each spot intensity (Freudenberg 2005). Thus, if m_j is a statistical measure on chip j that is equalised across the chips after normalisation, the scaling factor α_j on chip j is then given by

$$\alpha_j = \frac{m}{m_j}$$

   where m is the final value of m_j after normalisation.
2. *Transformation methods* rely on assumptions that allow quantitative mapping of two sets of spot intensities. The most popular methods are curve fitting, LOWESS and quantile normalisation. The curve fitting method assumes that the distribution of the normalised data set is known, and attempts to identify parameters of the distribution model; for instance, Lu et al. (2005) suggested adapting Zipf's law for the normalisation of one or two-colour DNA arrays (Draghici 2003; Lu et al. 2005). The LOWESS (locally weighted polynomial regression) method maps data from two data sets using a polynomial regression within overlapping intervals (Draghici 2003) and is most effective when most of the spots within two arrays show similar intensities (Oshlack et al. 2007). The quantile normalisation method can be applied where the assumption of a common underlying distribution seems to be justified. It is fast, easily implementable and does not require statistical modelling of the data (Freudenberg 2005; Bolstad et al. 2003).
3. *Invariant set method* relies essentially on the ability to identify a suitable set of non-differentially expressed probes or 'housekeeping' probes. The selection of the set of invariant spots might be experiment dependent, and an inappropriate choice of housekeeping probes can lead to bias in results (Freudenberg 2005; Ploner et al. 2005).

3.6.2 Protein Microarray Normalisation Methods

Custom antigen arrays however are typically not directly amenable to any of these standard DNA microarray normalisation approaches (Oshlack et al. 2007) since often a relatively small selection of specific probes show strong signals for a given sample and the identity of these probes vary between samples. This breaks most of the usual assumptions based on the comparison of equivalent signals across arrays,

unless the samples being compared display special features (Oshlack et al. 2007). Therefore, two methods are widely used for normalisation of antigen microarray data: the first relies on housekeeping controls and the second on a microarray sample pool (MSP) control (Lu et al. 2005; Oshlack et al. 2007; Wilson et al. 2003). Housekeeping controls are ideally supposed to keep a consistent signal across experimental conditions and different samples. In regard to antigen arrays, such as our CT100 array, used in assessing antibody profiles in serum, the housekeeping controls should ideally be serum independent to enable comparisons across different samples. However, identification of suitable housekeeping proteins in serum is less straightforward, as shown by the many mass spectrometry-based proteomic experiments that have documented high variability in the serum proteome; this housekeeping control-based normalisation approach therefore may be of limited value in serological assays using antigen microarrays.

Alternatively, the MSP method selects probes from a heterogeneous pool library and dilutes them at different concentrations to cover ranges similar to those covered by the probes used in the experiment; these probes must be printed in a number large enough to enable the assumption of non-differential expression between samples (Oshlack et al. 2007). Transformation methods such as LOWESS can subsequently be applied for normalisation. However, the limited size of the typical custom antigen microarrays can be a limiting factor in the usage of the MSP method and may restrict its readily application in serological assays using custom antigen microarrays.

3.6.3 Composite Normalisation Methods for Custom Antigen Microarrays

The normalisation method used by our group aims to make more efficient, effective and robust usage of a relatively small number of positive controls to correct for systematic bias in pin-to-pin and array-to-array variations. Robustness here is taken to mean the ability of the method to cope with the flagging of some positive controls, while still being based on sound biological principles.

The normalisation assumption we make here is that our positive control spots share a common underlying distribution across the chips (block, arrays, etc.) on which they are printed. This perspective provides greater flexibility than assuming that the individual positive control spots maintain the same intensities across the chips. Thereafter, our composite normalisation method combines quantile and total intensity normalisation modules to correct for systematic bias among the chips, while providing more robustness when dealing with flagged positive control spots (Causton et al. 2004; Bolstad et al. 2003).

1. *Quantile-Based Module*
   Since the positive control spots – in our CT100 array, these are Cy5-labelled, biotinylated BSA – are replica spots across different arrays, it seems reasonable to assume that they share an underlying distribution across arrays, and the quantile approach can be used to identify the corresponding housekeeping spot intensities based on their intensity distributions.

Bolstad et al. (2003) described an algorithm to carry out spot identification within the same quantile according to the following steps, where S_{ij} is the intensity of a positive control spot i on chip j:

(a) Load the positive control spot intensities S_{ij} into an $I \times J$ matrix X.
(b) Sort the spot intensities in each column j of X to get X_{sort}.
(c) Take the means across each row i of X_{sort} and get $\overline{X}_i$.

$\overline{X}_i$ is considered the underlying distribution of the positive control spot intensities across chips (Bolstad et al. 2003). This reorganisation enables more flexibility in handling outliers or flagged spots within the positive control data set.

2. *Total Intensity-Based Module*
This module assumes that post-normalisation, all arrays have a common total intensity value of their positive control spots (i.e. the sum of all the positive control spot intensities on each array should be constant) (Causton et al. 2004; Quackenbush 2001), given by $\sum_{i=1}^{N_{\text{spots}}} \overline{X}_i$. The normalisation factor α_k to normalise array k is then given by

$$\alpha_k = \frac{\sum_{i=1}^{N_{\text{spots}}} \overline{X}_i}{\sum_{i=1}^{N_{\text{spots}}} \overline{X}_{ik}}$$

where $\sum_{i=1}^{N_{\text{spots}}} \overline{X}_{ik}$ is the total intensity of all the positive control spots on array k prior to normalisation.

This is a scaling normalisation method that assumes that different arrays share a common total intensity of their housekeeping spots, while taking into account the potential existence of flagged spots within the housekeeping spots. Importantly, if a given positive control spot is identified as an outlier on one array (i.e. it is flagged for some reason), the corresponding positive control spots across all arrays are identified in the quantile module above and are then also flagged across all arrays prior to normalisation; the net consequence of this is to ensure that the same number of positive control spots are considered across all arrays during normalisation.

3.6.4 Overall Preprocessing Pipeline

Our overall workflow, illustrated in Fig. 3.5, includes the following steps:

1. Spot filtering flags spots whose fluorescent pixels comprise less than 20% of an arbitrarily defined spot area, as well as spots that show saturation in >10% of the pixels.
2. Background adjustment corrects the local background of each spot to become the median neighbourhood background of its surrounding spots (3×3 window).

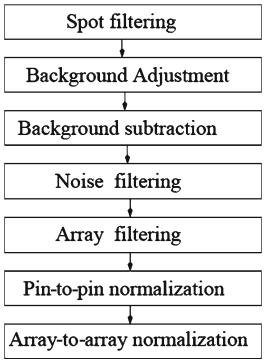

Fig. 3.5 Schematic of overall antigen microarray data preprocessing pipeline

3. Background subtraction gives the net spot intensity by subtracting the corrected neighbourhood background from the original raw median pixel intensities.
4. Noise filtration sets all net intensities to zero if they are lower than 2 SD of the background.
5. Array filtering calculates the CV of the positive controls within the array prior to normalisation and flags arrays with CVs above a user-defined threshold (we typically use a value of 30% here).
6. Pin-to-pin normalisation normalises the data based on our composite normalisation method in order to account for variations between the usage of different pins during the print run.
7. Array-to-array normalisation normalises the data, again using our composite method, in order to account for variations between arrays and to thereby allow, for example, comparison of serology data obtained on samples collected at different time points from the same patient (Safari Serufuri 2010).

3.7 Protein Microarray Data: Qualitative Clustering

Following preprocessing, quality control and normalisation of antigen microarray data, quantitative bioinformatic methods can then be applied with greater confidence to enable biological interpretation of the data.

The ultimate purpose of all clustering methods is to group or segment a set of items by taking into consideration a criterion of similarity or dissimilarity. The clustering can provide information regarding data structures as well as outliers – an outlier being an item not sufficiently similar to any other items in the data set. An additional goal of clustering can be to infer hierarchical order between clusters (Draghici 2003; Causton et al. 2004; Hastie et al. 2001; Boutros and Okey 2005; Costello and Osborne 2005).

Before choosing a clustering algorithm, one has to determine the intended type of cluster that might be expected from the data set and the most appropriate measure of similarity to capture the clusters of interest. Another essential consideration

is the performance of the selected algorithm (Hastie et al. 2001; Boutros and Okey 2005). Qualitative clustering can be based on the trend line similarities between antibody profiles of patient samples at a common time point or on trend line similarities between patients across a timecourse (e.g. trend line similarities in changes in autoantibody profiles between samples collected at different time points posttreatment).

We routinely use two clustering methods for analysis of our CT antigen array data: a factor analysis method and a *K*-means method using a Pearson correlation metric. Detailed mathematical description of these two clustering methods is beyond the scope of this chapter but a brief qualitative description is as follows:

The factor analysis method – used in an unsupervised mode – aims to investigate the number of intrinsic factors that are required to account for the correlations among variables or observations (Hastie et al. 2001). Factor analysis has been widely used in intelligence research to explain a variety of results based on different tests by identifying groups of correlated results. For instance, the performance at running, weight lifting and jumping could be explained by general athletic ability (Costello and Osborne 2005; Wikipedia 2010; Tryfos 2010).

The *K*-means method is among the most popular clustering algorithms. It is an iterative method relying on the minimisation of an objective function defining a measure of dissimilarity between the items or of their K-centroids, a centroid here being a measure by which the dissimilarity between clusters or between clusters and items can be summarised by one value (Causton et al. 2004; Hastie et al. 2001; Boutros and Okey 2005). The *K*-means method however requires a number of preconceived inputs to achieve the clustering, including the number of expected clusters K, the maximum number of iterations, the selection of the similarity metric (Pearson, Euclidian, etc.) and an initial clustering which is improved iteratively until a steady state or the maximum number of iterations is reached (Draghici 2003).

3.8 Experimental Design: Test Case of a Cancer–Testis Antigen Array for In Vitro Cancer Biomarker Discovery

We have applied our antigen microarray approach in cancer biomarker discovery projects and describe below a test case using serum samples from patients undergoing an experimental cancer treatment, the aim being to use our CT100 microarray to monitor changes in the autoimmune profiles of those patients following treatment as a possible correlate of therapeutic response.

3.8.1 Description

The ‘CT100’ array is a ‘one-colour’ CT antigen microarray designed to discover the expression profile of 100 CT- or tumour-associated (TA) antigens of interest (see Table 3.1) in specific cancer patients, as revealed by the binding of autoantibodies

Table 3.1 List of the 72 CT antigens (*yellow*) and the 28 non-CT antigens of interest (*blue*) present within each array field and comprising the 'CT100' array

Antigen identity	Antigen identity	Antigen identity	Antigen identity
BAGE2	LEMD1	SGY-1	CDK7
BAGE3	LIPI	SILV	FES
BAGE4	MAGEA1	SPAG9	FGFR2
BAGE5	MAGEA10	SPANXA1	MAPK1
CCDC33	MAGEA11	SPANXB1	MAPK3
CEP290	MAGEA2	SPANXC	PRKCZ
COL6A1	MAGEA3	SPANXD	RAF
COX6B2	MAGEA4 v2	SPO11	SRC
CSAG2	MAGEA v3	SSX1	CALM1
CT47.11	MAGEA v4	SSX2a	CDC25A
CT62	MAGEA5	SSX4	CREB1
CTAG2	MAGEB1	SYCE1	CTNNB1
CXorf48.1	MAGEB5	SYCP1	p53 S6A
DDX53	MAGEB6	THEG	p53 C141Y
MMA1	MART-1/MLANA	TPTE	p53 S15A
FTHL17	MICA	TSGA10	p53 T18A
GAGE1	NLRP4	TSSK6	p53 Q136X
GAGE2A	NXF2	TYR	p53 S46A
GAGE4	NY-CO-45	XAGE-2	p53 K382R
GAGE5	NY-ESO-1	XAGE3a v1	p53 S392A
GAGE6	OIP5	XAGE3a v2	p53 M133T
GAGE7	p53	ZNF165	p53 L344P
GRWD1	PBK	AKT1	CYP3A4
HORMAD1	RELT	CDK2	CYPR
LDHC	ROPN1	CDK4	EGFR

present in a patient's serum to the purified and spatially segregated, immobilised antigens. Uses of this CT antigen array include monitoring therapeutic responses and/or the rate of cancer progression in individual patients (Safari Serufuri 2010).

Recombinant gene cloning methods were used to clone each antigen into a relevant insect cell expression vector as a C-terminal fusion construct with the biotin carboxyl carrier protein (BCCP) tag (see Supplementary Material for detailed protocol). The BCCP tag is biotinylated in vivo in insect cells (as well as in *E. coli* and yeast) and allows single-step purification and immobilisation onto the array surface via the high specificity and affinity streptavidin–biotin interaction. Expression and biotinylation of each recombinant antigen in insect cells (see Supplementary Material for detailed protocol) was confirmed using Western blot analysis prior to printing, and crude lysates were then diluted with PBS containing 40% sucrose. Replica 'CT100' protein arrays were printed in a 4-plex format (i.e. 4 replica fields per slide) using crude cell lysates of Sf21 insect cells expressing each of the 72 CT antigens and 28 TA antigens of interest. Various controls were also included in each array field: 50 ng/μl human

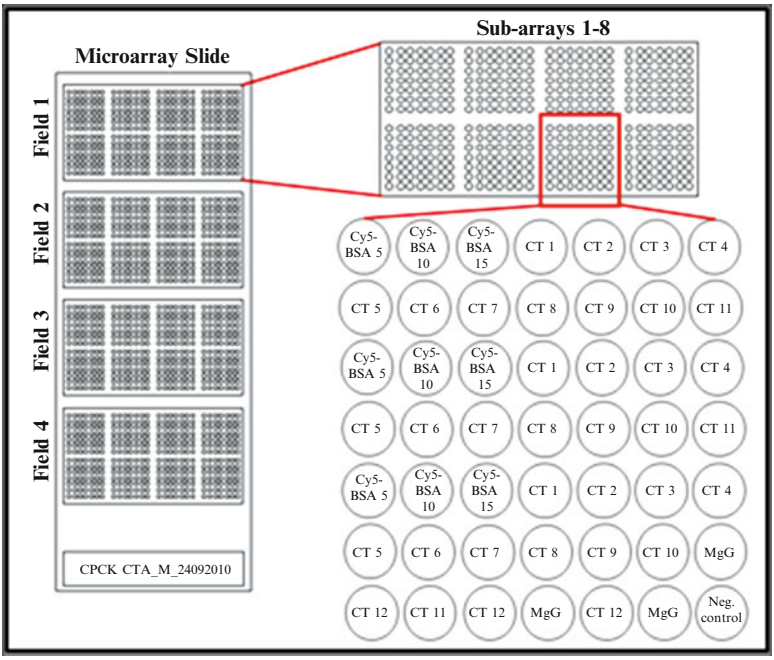

Fig. 3.6 Schematic of the CT100 array layout. Each slide contains four replicate fields each comprising all 100 antigens of interest (one field is used per patient sample assay). Each field contains eight blocks of 7×7 spots each. A representative block is shown with the *Cy5-BSA* controls, the various antigens, the *hIgG* positive control and a negative control

IgG (positive control), 200 ng/µl sheep IgG (negative control), as well as a crude insect cell lysate expressing BCCP only and no fusion protein (negative control). For slide orientation and signal normalisation, three different biotinylated Cy5-BSA concentrations were included in each sub-array (5, 10 and 15 ng/µl). Each sample and control was spotted in triplicate within each array field, while the Cy5-BSA concentration series was spotted in triplicate by each pin within each sub-array (i.e. 24 spots total per array field). Array design is shown in Fig. 3.6 below (Beeton-Kempen et al. submitted).

Each CT100 array was printed on an in-house streptavidin-coated surface (Nexterion slide H) using a Genetix QArray2 (Genetix Ltd., UK) robotic microarrayer with 8×300 µm flat-tipped solid pins (see Supplementary Material for detailed protocol). After printing, each slide was washed with prechilled blocking solution (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 1 mM DTT and 50 µM biotin) and stored at –20°C in storage buffer (same as the blocking solution except with 50% glycerol and no biotin). Under these conditions it proved possible to store these slides for up to 3 weeks prior to assay (Beeton-Kempen et al. submitted).

3.8.2 CT100 Assay Using Patient Serum

A collection of 100 serum samples from a malignant melanoma patient cohort (UCT HREC Ref number: 240/2011) was subjected to serological assay as follows:

Each serum sample was diluted 1:800 in PBS/0.1% Tween-20 and incubated on an individual CT100 array (prepared as described above, Sect. 3.8.1) at room temperature for 1 h. Each array was then washed independently with slide buffer (PBS/0.1% Tween-20). Cy5-labelled goat antihuman IgG (Invitrogen) was diluted 1:100 in slide buffer and incubated with each individual array for 1 h at RT. The individual arrays were then washed independently in slide buffer, dried and then scanned immediately using a Tecan LS Reloaded fluorescence microarray scanner (Tecan Group Ltd., Switzerland). All liquid handling steps were carried out on a Tecan HS4800 Pro automated hybridisation station (Beeton-Kempen et al. submitted). The arrays were scanned at a resolution of 20 μm using fixed gain settings of 110, 120, 125 and 135 PMT in order to determine the setting that gave the highest signal with minimal saturation across all slides.

3.8.3 CT100 Array Data Extraction

Using ArrayPro Analyzer v6.3 software (Media Cybernetics Inc., USA), a grid was semiautomatically autoaligned over the individual spots on each image file. A constant area feature finder was used across the array surface. ArrayPro was then used to extract the raw data from each of the spots (in batch processing mode), using a 200-RFU pixel threshold and a .gal file (output from the Genetix QArray2 printer) containing information about the identity of the sample in each spot. Upon extraction, ArrayPro provided both the mean and median foreground and background pixel intensities of each spot. This data was then used for further processing and analysis (Beeton-Kempen et al. submitted).

3.8.4 CT100 Array Data Processing

The net intensity for each spot was calculated by subtracting the corrected neighbourhood median background of surrounding spots from the local median foreground intensity of each spot. The average of the three replicate spots' net intensities was calculated to determine the mean signal for each sample. The data was then filtered to remove all spots displaying saturated signals, signals below the background signal plus two standard deviations noise threshold or where the signals observed occupied less than 20% of the area of the captured spot. Each array (corresponding to a unique patient serum sample) was inspected, and all array fields displaying >30% coefficient of variation across the Cy5-BSA spots were excluded from further analysis and the corresponding samples re-assayed on new arrays.

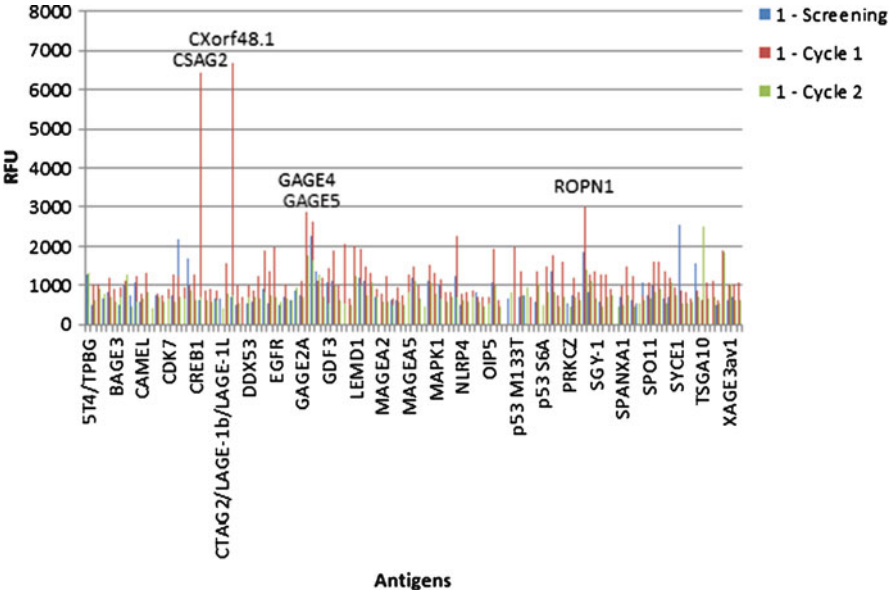

Fig. 3.7 Graph displaying the autoimmune profile of Patient 1 pre and post experimental treatment

Array-to-array normalisation was carried out using the net signal intensities of the 15 ng/µl biotinylated Cy5-BSA spots of each array and across each slide using the composite normalisation method developed by our group, as described above (Sect. 3.6.3 & 3.6.4) (Safari Serufuri 2010). All array images were also visually inspected for evidence of spot bleeding, washing artefacts or pin sticking, and relevant arrays were excluded from further analysis and the corresponding samples re-assayed on new arrays.

3.8.5 CT100 Array Results and Discussion

Serum samples were taken from patients prior to receiving an experimental treatment ('screening') and then following different cycles of treatment (cycles 1, 2, etc.). Not all patients had the same number of treatment cycles nor were these all carried out at the same time intervals. Nevertheless, distinct anti-CT antigen autoimmune responses were observed for the majority of this melanoma cohort prior to treatment, and these patterns appeared to be modified significantly in response to treatment (see, e.g. Figs. 3.7 and 3.8). Note that these two patients showed very different autoimmune profiles prior to treatment as well as differential responses to the experimental treatment. Here, erring on the side of caution, we interpret antigen signals from the array with net intensity >1,000 RFU as real and significant data. When comparing different

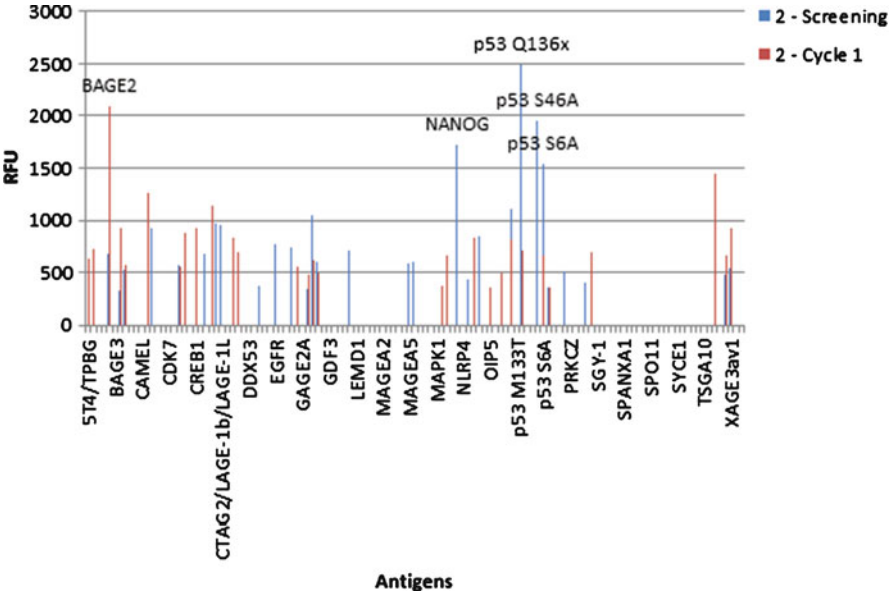

Fig. 3.8 Graph displaying the autoimmune profile of Patient 2 pre and post experimental treatment

time points for a given patient, simple statistical calculations (e.g. *t*-tests or ANOVA) – based on the triplicate repeat data for each antigen on each array – can be performed to determine whether the intensity at one time point is significantly different to that at a later time point. As expected, the number, identity and strength of anti-CT antigen autoimmune responses varied from patient to patient across our cohort, as exemplified in Figs. 3.7 and 3.8. In addition, when comparing the autoimmune profile of Patient 1–2, it is evident that Patient 1 has a noisier array, with nearly all antigens lighting up between 0 and 500 RFU, while Patient 2 has a cleaner array, with approximately only 30 antigens of interest showing strong signals above 500 RFU. The biological rationale underlying this observation remains unclear as yet, but it is noteworthy that all signals shown in Figs. 3.7 and 3.8 are significant compared to background. It is also noteworthy that we have previously verified individual anti-CT antigen responses by Western blot and have also compared our microarray data to ELISA data for individual antigens where available: in all cases, such comparisons have provided independent verification of the specificity of our protein microarray data (data not shown) (Beeton-Kempen et al. submitted). Furthermore, we have also demonstrated that our CT100 microarray platform shows linearity of response to anti-CT antigen autoantibody titres across 3–4 orders of magnitude and that it has a limit of quantitation in the range of 100 pg/ml (data not shown) (Beeton-Kempen et al. submitted).

The freely available *MultiExperiment Viewer* (MeV; v4.8.1) software was used to cluster the above data using the *K*-means algorithm using the Pearson correlation

(see Fig. 3.9). N.B. MEV was utilised solely for clustering purposes and not for normalisation or other data adjustments.

The resultant heatmap (Fig. 3.9) provides a compact visualisation of the data, which facilitates data interpretation and comparison between patients.

3.9 Conclusions

In this chapter we have highlighted the importance of both experimental and informatic protocols to minimise the influence of variations within microarray data sets due to non-specific binding, smears, artefacts or high background resulting from ineffective washing of the array surface, any of which can confound interpretation of data from a protein microarray experiment. Standardising protein microarray workflows and methodologies is essential to allow comparison of data generated at different times by the same or different laboratories.

We have emphasised the deterministic role of an experimental design and suggested the importance of the selection of specific controls, to enable normalisation of data and to assess background binding effects. With the development of appropriate preprocessing and quality control pipelines for raw array data, qualitative clustering of patient samples based on their autoantibody profiles measured on such antigen arrays becomes viable. We note that characterisation of anti-CT antigen autoantibody profiles from a single patient provides qualitative data on CT antigen expression, while comparison of anti-CT antigen autoantibody profiles across a timecourse for a given patient enables quantitative data to be generated on changes in autoantibody titres (Casiano et al. 2006). Factor analysis allows an unsupervised approach to cluster samples and provides a qualitative measure of how profiles correlate to a given cluster while also enabling the detection of outlier profiles within clusters. Compared to the *K*-means method, factor analysis is more straightforward, since it directly identifies the number of potential clusters and does not proceed by iterative steps. Factor analysis also provides more reproducible results, which is not always the case with *K*-means clustering, where final clustering might depend on the initial cluster assignment – which is randomly generated – and a sufficient number of iterations. In our experience, the limited number of control spots typically printed onto custom antigen arrays proved to be a challenge when determining the appropriate controls to generate a robust normalisation hypothesis, particularly since serum compositions are known to be strongly patient dependent. Therefore, the future utility of custom antigen microarrays in cancer biomarker discovery research will rely strongly on their design, which is ultimately linked to the available knowledge on the research question being addressed in each particular experiment. Despite these challenges though, the fast evolving protein microarray field has already enabled the identification of several serum antigens and antibody markers in cancer, although it is worth noting that validation of such candidate cancer biomarkers has in general yet to be confirmed (Matarraz et al. 2011).

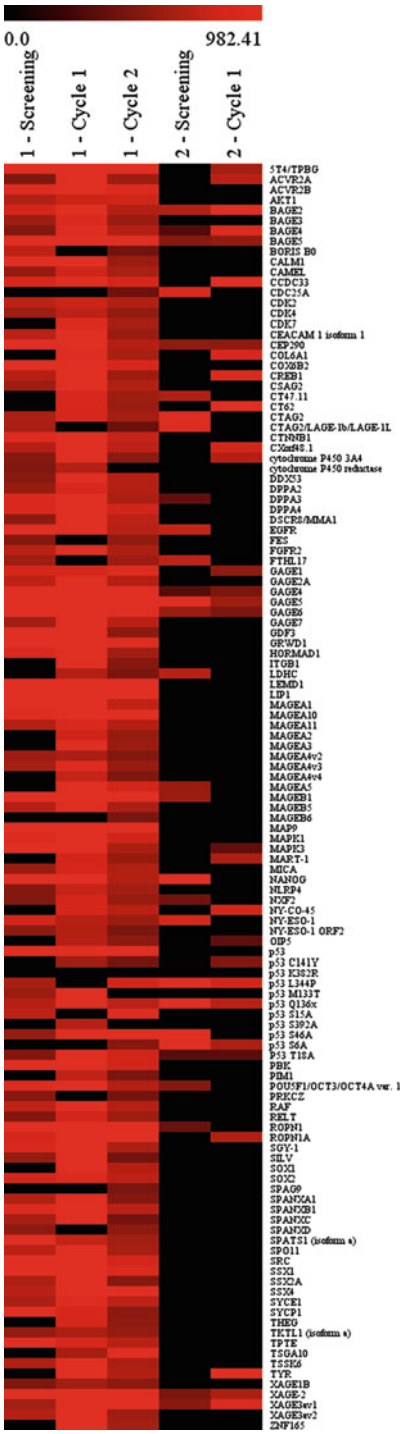

Fig. 3.9 Heatmap of Patient 1 and 2 autoimmune profiles at two time points following *K*-means clustering

Acknowledgements The authors thank Dr Aubrey Shoko, Dr Natasha Beeton-Kempen and Dr Judit Kumuthini for their help in generating the data herein. We thank the Centre for Proteomic & Genomic Research, Cape Town, for access to equipment and assistance in developing the CT100 array. JMB thanks the National Research Foundation (NRF), South Africa, for a Research Chair. The research was supported by grants from the NRF, University of Cape Town (UCT) and Marion Beatrice Waddel.

Supplementary Material

Methodology

Cloning of Cancer/Testis Antigen Genes

In total, 100 proteins were cloned and expressed for printing on the CT100 array. Seventy-two of these were CT antigen proteins, while the remaining twenty-eight were other cancer-associated proteins/proteins of interest. All antigens were cloned into baculoviral expression vectors and expressed in insect cells.

The following procedure was carried out for insect cell-expressed proteins. The gene encoding the *E. coli* biotin carboxyl carrier protein (BCCP) domain – amino acids 74–156 of the *E. coli accB* gene (Athappilly and Hendrickson 1995; Chapman-Smith and Cronan 1999) – was amplified by PCR from an *E. coli* genomic DNA preparation and cloned downstream of a viral polyhedrin promoter in an E. coli vector to create the transfer vector pJB1. This *E. coli* transfer vector system is derived from pTriEx-1.1 (Novagen). Flanking this *polh*-BCCP expression cassette were the baculoviral 603 and 1,629 genes (Zhao et al. 2003), which enabled the subsequent homologous recombination of the construct into a replication-deficient baculoviral genome.

Synthetic genes for each of the antigens of interest were obtained from Origene, Open Biosystems or GeneService. PCR primers were designed for each CT antigen cDNA such that the stop codon would be removed, enabling it to be cloned into the pJB1 transfer vector upstream of and in-frame with the 3′-BCCP tag via ligation-independent cloning methods, replacing the ORF region between the *Spe* I and *Nco* I sites of pJB1 in the process (all primers synthesised by IDT, UK) (Yang et al. 1993). Each resulting transfer vector thus encoded an individual antigen fused to a C-terminal BCCP tag.

The PCR amplification of each synthetic gene; ligation-independent cloning of these products into a BCCP tag-containing transfer vector, pJB30; and transformation of this vector into *E. coli* DH5α were all carried out according to standard recombinant DNA protocols (Sambrook et al. 2001) and are accordingly not described here in detail. Successful PCR amplification of each antigen was confirmed by gel electrophoresis, while successful cloning was determined by sequencing the relevant regions of each transfer vector (i.e. the region containing the ligated PCR

products as well as the junctions between these and the BCCP tags) according to standard protocols and verified against the RefSeq database.

Maintenance and Co-transfection of Sf21 Cells

A replication-incompetent baculovirus vector, bacmid pBAC10:KO_{1629} (Zhao et al. 2003), was propagated in *E. coli* HS996 cells, and bacmid DNA was prepared according to standard procedures. pBAC10/KO_{1629} was then linearised by restriction with *Bsu*361 (New England Biolabs) for 5 h at 37°C, after which *Bsu*361 was heat killed at 80°C for 15 min. Five hundred nanograms of undigested pJB1 transfer vector was then combined with 500 ng linearised bacmid, and the total volume made up to 12 μl with water. Twelve microlitres Lipofectin (diluted 2:1 in H_2O) was then added to this DNA mix, and the tube was incubated at room temperature for 30 min. One millilitre serum-free media (InsectXpress, Lonza) was added to the Lipofectin/DNA mixture. A 6-well plate containing 1×10^6 Sf21 cells/well (Invitrogen) was prepared and incubated at 27°C for 1 h to allow the cells to adhere. Excess media were aspirated from the Sf21 cells and replaced with the Lipofectin/DNA/serum-free mix. The transfected cells were incubated at 27°C overnight. The media was then replaced with 2 ml InsectXpress media supplemented with 2% FBS and incubated at 27°C without agitation for a further 72 h. Cells were resuspended by physical agitation and then pelleted by centrifugation at 1,000 × *g* for 10 min. The supernatant containing recombinant baculovirus was transferred to a fresh tube and stored at 4°C; this was the P_0 stock. The general baculoviral system used here was adapted from the work of Prof Ian Jones (Reading University, UK) (Zhao et al. 2003).

Recombinant baculoviral particles were amplified according to standard procedures. Briefly, a 6-well plate was set up with 1×10^6 Sf21 cells/well and incubated at 27°C for 1 h. Excess media were removed and replaced with 500 μl of P_0 virus plus 500 μl InsectXpress media supplemented with 2% FBS and incubated at 27°C without agitation for a further 72 h. P_1 virus was harvested as described above. A 150-ml tissue culture flask was seeded with 20 ml of 1×10^6 Sf21 cells/ml and incubated at 27°C for 1 h. Excess media were removed and replaced with 500 μl of P_1 virus plus 3 ml InsectXpress media supplemented with 2% FBS and incubated at 27°C for 1 h, after which a further 25 ml InsectXpress media supplemented with 2% FBS were added and cells incubated without agitation for 72 h. P_2 virus was harvested as described above. The titre of the P_2 viral stock was determined by a SybrGreen-based quantitative PCR assay versus a stock of known titre determined by plaque assay. Stocks that were found to have low titre were re-amplified.

Expression of BCCP-Tagged CT Antigens

A 24-well deep well plate containing 6×10^6 Sf21 cells/well suspended in 3 ml InsectXpress media supplemented with 2% FBS and 50 μM biotin was used. 200 μl

of P_2 virus was added, and the plate was incubated at 27°C for 72 h with agitation. Cells were harvested by centrifugation of the plate prior to lysis. Cells were gently resuspended and washed in 3 ml of PBS buffer for 5 min, the plate was recentrifuged and the supernatant was discarded; this was repeated three times in total. Pellets were gently resuspended in 350 µl of freezing buffer (25 mM HEPES, 50 mM KCl, pH 7.5) ensuring thorough mixing of the cells. Cells were aliquoted in 50 µl volumes and stored at −80°C until required for cell lysis. For cell lysis, aliquots were thawed and 50 µl lysis buffer (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 250 U/ml protease inhibitor cocktail and 1 mM DTT plus 10 U Benzonase (Novagen)) was added to each; this was then incubated on ice with agitation for 30 min. Cell debris was removed by centrifugation at 13,000 × *g* for 30 min at 4°C, the supernatant collected and then stored on ice for up to 24 h prior to printing.

The protein concentration of the soluble, crude protein extract was determined by Bradford assay (Bradford 1976) to confirm that effective cell lysis had occurred. Antigen expression and biotinylation were analysed by Western blot according to standard protocols (Sambrook et al. 2001). Antigen expression was confirmed using a mouse anti-c-myc antibody (Sigma-Aldrich) at 1:5,000 followed by a 1:25,000 dilution of goat anti-mouse IgG HRP conjugate (KPL). For more rapid processing, dot blots were sometimes used (same conditions as for Western blots) to assess expression prior to array fabrication. Biotinylation of the antigens was confirmed using a streptavidin–HRP conjugate probe (GE Healthcare) at 1:10,000.

Fabrication of Protein Microarrays

Preparation of Streptavidin-Coated Slides for Printing

A Nexterion Slide H microarray slide (Schott, Germany) was equilibrated to room temperature and removed from the foil package. A 1 mg/ml streptavidin solution was made up in 150 mM of Na_2HPO_4 buffer (pH 8.5). The microarray surface was immersed in approximately 5 ml of the streptavidin solution for 1 h at room temperature. The slide was removed from the streptavidin solution (which can be reusable successively up to 10 times) and then washed for 1 h at room temperature in 10 ml of 150 mM Na_2HPO_4 buffer (pH 8.5) containing 50 mM of ethanolamine to deactivate any remaining amine-reactive groups. The slide was washed for 3 × 5 min in 10 ml wash buffer and then for 5 min in 10 ml water. The slide was then placed in a 50-ml Falcon tube and centrifuged at 1,000 × *g* for 5 min at 20°C until dry. Streptavidin-coated slides were placed into slide boxes, sealed in Ziploc bags and stored at −20°C.

As a QC test, one streptavidin-coated slide per batch was incubated for 1 h with a solution of Cy5-biotinylated BSA (10 µg/ml in PBS), washed and scanned; this demonstrated that with this procedure, we can readily achieve CVs of 2–3% across the print area of the slide surface, judged by analysis of a virtual grid of 576 evenly distributed spots.

CT Antigen Microarray Fabrication

The expression and biotinylation of the various antigens were confirmed using SDS-PAGE- or dot blot based Western blot analysis prior to printing and crude lysates were then diluted with PBS containing 40% sucrose (sucrose was included to increase the surface tension and to reduce spreading of printed droplets). Forty microlitres of the crude protein extract for each BCCP-tagged protein to be arrayed was transferred into individual wells of a 384-well V-bottom plate. The plate was centrifuged at 4,000 rpm for 2 min at 4°C to pellet any cell debris that may have carried over from cell lysis. The plate was then stored on ice prior to the microarray print run, and during printing it was kept at 4°C.

Replica CT100 arrays were printed in a 4-plex format (i.e. 4 replica arrays per slide), using crude cell lysates. Each of the 72 CT antigens and the 28 TA antigens were printed in triplicate within each array. Several different controls were also included in each array. The positive controls included 50 ng/μl biotinylated human IgG (Rocklands Immunochemicals Inc.). The negative controls included biotinylated 200 ng/μl sheep IgG (Rocklands Immunochemicals Inc.) and an 'empty vector' lysate control consisting of a crude insect cell lysate containing the BCCP-tag alone with no recombinant fusion partner. In addition, three different concentrations (5, 10 and 15 ng/μl) of biotinylated Cy5-BSA were included in each sub-array for slide orientation and signal normalization purposes.

Each CT100 array was printed on home-made streptavidin-coated microarray slides (prepared as above) using a Genetix QArray2 robotic arrayer (Genetix Ltd., UK) equipped with 8 × 300 μm flat-tipped solid pins. Each array was printed as a set of eight 7 × 7 blocks, with each block printed by a different pin. The printing procedures were carried out at room temperature, while the source plate was kept at 4°C, and the atmosphere in the print chamber was humidified to ~50%. The arrays were printed using the following key QArray2 settings: inking time = 500 ms, microarraying pattern = 7 × 7, 500 μm spacing, maximum stamps per ink = 1, number of stamps per spot = 2, printing depth = 150 μm, water washes = 60 s wash and 0 s dry, ethanol wash = 10 s wash and 1 s dry.

After printing, each slide was washed for 30 min with 50 ml prechilled blocking solution (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 1 mM DTT and 50 μM biotin) and then stored at −20°C submerged in storage buffer (25 mM HEPES pH 7.5, 50% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA and 1 mM DTT).

Verification of Immobilisation of BCCP-Tagged Proteins to Array Surface

Following standard protocols for Western blots, it is possible to verify the successful immobilisation of biotinylated proteins to the array surfaces, as follows. Mouse anti-c-myc antibody was diluted 1:1,000 in 1 ml PBST containing 5% fat-free milk powder. The protein array was removed from wash buffer and equilibrated in PBST at room temperature for 5 min. The PBST was drained away and 5 ml antibody solution

was added to the array, which was then incubated with gentle agitation at room temperature for 30 min. The array was washed for 3×5 min with 1 ml of PBST. Goat anti-mouse antibody-HRP conjugate was diluted 1:1,000 in 1 ml milk/PBST. The antibody solution was added to the array, and the array was incubated with gentle agitation at room temperature for 30 min. The array was washed for 3×5 min with 1 ml of PBST and then submerged in 5 ml of chemiluminescent detection reagents (Pierce). After 1 min, the slide was placed in a 50-ml Falcon tube and centrifuged for 30 s to dry. In a dark room, the array was placed against autoradiography film for varying lengths of time before developing the film.

References

Altman N. Replication, variation and normalization in microarray experiments. Appl Bioinformatics. 2005;4:1–23.

Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res. 2005;4:1123–33.

Angenendt P, Glökler J, Sobek J, Lehrach H, Cahill DJ. Next generation of protein microarray support materials: evaluation for protein and antibody microarray applications. J Chromatogr. 2003;1009:97–104.

Athappilly FK, Hendrickson WA. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. Structure. 1995;3:1407–19.

Beeton-Kempen N, Duarte JG, Shoko A, Safari Serufuri J-M, Cebon J, Blackburn JM. Monitoring melanoma patient responses to therapeutic vaccination using a cancer/testis antigen protein microarray. Manuscript submitted.

Berrade L, Garcia AE, Camarero JA. Protein microarrays: novel developments and applications. Pharm Res. 2011;28:1480–99.

Blackburn JM, Shoko A. Protein function microarrays for customised systems-oriented proteomic analysis. In: Korf U, editor. Protein microarrays: methods and protocols, Methods in molecular biology. Springer protocols. New York: Humana Press; 2011. Chapter 21. ISBN 978-1-61779-285-4.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.

Boutell JM, Hart DJ, Godber BLJ, Kozlowski RZ, Blackburn JM. Functional protein microarrays for parallel characterisation of p53 mutants. Proteomics. 2004;4:1950–8.

Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005;6:331–43.

Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976;72:248–54.

Brusic V, Marina O, Wu CJ, Reinherz EL. Proteome informatics for cancer research: from molecules to clinic. Proteomics. 2007;7:976–91.

Büssow K, Konthur Z, Lueking A, Lehrach H, Walter G. Protein array technology: potential use in medical diagnostics. Am J Pharmacogenomics. 2001;1:1–7.

Casiano CA, Mediavilla-Varela M, Tan EM. Tumor-associated antigen arrays for the serological diagnosis of cancer. Mol Cell Proteomics. 2006;5:1745–59.

Causton HC, Quackenbush J, Brazma A. Microarray gene expression data analysis: a beginners guide. 1st ed. Malden: Blackwell Publishing; 2004.

Chapman-Smith A, Cronan JE. The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity. Trends Biochem Sci. 1999;24:359–63.

Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Prac Assess Res Eval. 2005;10:1–9.

Draghici S. Data analysis tools for DNA microarrays. 2nd ed. Boca Raton: Chapman & Hall; 2003.
Espina V, Mehta AI, Winters ME, Calvert V, Wulfkuhle J, Petricoin III EF, et al. Protein microarrays: molecular profiling technologies for clinical specimens. Proteomics. 2003;3: 2091–100.
Fang Y, Lahiri J, Picard L. G protein-coupled receptor microarrays for drug discovery. Drug Discov Today. 2003;8:755–61.
Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev. 2003;2:566–80.
Freudenberg JM. Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. Leipzig Bioinformatics Working Paper. 2005;3:1–120.
Gray MR, Colot HV, Guarente L, Rosbash M. Open reading frame cloning: identification, cloning, and expression of open reading frame DNA. Proc Natl Acad Sci U S A. 1982;79:6598–602.
Hall DA, Ptacek J, Snyder M. Protein microarray technology. Mech Ageing Dev. 2007;128:161–7.
Hardiman G. Microarray technologies – an overview. Pharmacogenomics. 2003;4:251–6.
Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 1st ed. New York: Springer; 2001.
He M, Taussig MJ. Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). Nucleic Acids Res. 2001;29:73–3.
Hultschig C, Kreutzberger J, Seitz H, Konthur Z, Bussow K, Lehrach H. Recent advances of protein microarrays. Curr Opin Chem Biol. 2006;10:4–10.
Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:211–15.
Ingvarsson J, Larsson A, Sjo AG, Truedsson L, Jansson B, Borrebaeck CAK, et al. Design of recombinant antibody microarrays for serum protein profiling: targeting of complement proteins research articles. J Proteome Res. 2007;6:3527–36.
Klein JB, Thongboonkerd V. Overview of proteomics. In: Thongboonkerd V, Klein JB, editors. Proteomics in nephrology. Basel: Karger; 2004. p. 1–10.
Kodadek T. Protein microarrays: prospects and problems. Chem Biol. 2001;8:105–15.
Koopmann J-O, Blackburn J. High affinity capture surface for matrix-assisted laser desorption/ ionisation compatible protein microarrays. Rapid Commun Mass Spectrom. 2003;17:455–62.
Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
Lu T, Costello CM, Croucher PJP, Häsler R, Deuschl G, Schreiber S. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005;6:1–13.
Macbeath G. Protein microarrays and proteomics. Nat Genet. 2002;32:526–32.
MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. Science. 2000;289:1760–3.
Matarraz S, González-González M, Jara M, Orfao A, Fuentes M. New technologies in cancer. Protein microarrays for biomarker discovery. Clin Transl Oncol. 2011;13:156–61.
Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, et al. Analyzing antibody specificity with whole proteome microarrays. Nat Biotechnol. 2003;21:1509–12.
Oshlack A, Emslie D, Corcoran LM, Smyth GK. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. Genome Biol. 2007;8:2.1–8.
Phizicky E, Bastiaens PIH, Zhu H, Snyder M, Fields S. Protein analysis on a proteomic scale. Nature. 2003;422:208–15.
Ploner A, Miller LD, Hall P, Bergh J, Pawitan Y. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinformatics. 2005;6:1–20.
Predki PF. Functional protein microarrays: ripe for discovery. Curr Opin Chem Biol. 2004;8:8–13.
Quackenbush J. Computational analysis of microarray data. Genetics. 2001;2:418–27.
Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, Lau AY, et al. Self-assembling protein microarrays. Science. 2004;305:86–90.
Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24:971–83.
Robinson WH. Antigen arrays for antibody profiling. Curr Opin Chem Biol. 2006;10:67–72.

Safari Serufuri J-M. Development of computational methods for Custom protein arrays analysis. A case study on a 100 protein (“ CT100 ”) cancer/testis antigen array. Masters thesis, University of Cape Town. 2010.
Sambrook J, Russel DW, Macallum P. Molecular cloning – a laboratory manual. 3rd ed. Cold Spring Harbour: Cold Spring Harbour Laboratory Press; 2001.
Sanchez-Carbayo M. Antibody arrays: technical considerations and clinical applications in cancer. Clin Chem. 2006;52:1651–9.
Scanlan MJ, Gure AO, Old LJ, Chen Y-t. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev. 2002;188:22–32.
Schäferling M, Nagl S. Optical technologies for the read out and quality control of DNA and protein microarrays. Anal Bioanal Chem. 2006;385:500–17.
Schmidt DMZ, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, et al. Evolutionary potential of (b/a)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. Biochemistry. 2003;42:8387–93.
Schweitzer B, Predki P, Snyder M. Microarrays to characterize protein interactions on a whole-proteome scale. Proteomics. 2003;3:2190–9.
Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31:265–73.
Steinhoff C, Vingron M. Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform. 2006;7:166–77.
Tecan LS™ Series Laser Scanner: how to set the correct gain in the LS scanner. http://www.tecan.com
Tryfos P. Notes on Factor analysis. 2010. http://www.yorku.ca/ptryfos/f1400.pdf
Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature. 2003;423:506–11.
Wikipedia. Factor analysis in psychometrics. 2010. http://en.wikipedia.org/wiki/Factor_analysis
Wilson DL, Buckley MJ, Helliwell CA, Wilson IW. New normalization methods for cDNA microarray data. Bioinformatics. 2003;19:1325–32.
Wise E, Yew WS, Babbitt PC, Gerlt JA, Rayment I. Homologous (b/a)8 -barrel enzymes that catalyze unrelated reactions: orotidine 5′-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase. Biochemistry. 2002;41:3861–9.
Wolf-Yadlin A, Sevecka M, MacBeath G. Dissecting protein function and signaling using protein microarrays. Curr Opin Chem Biol. 2009;13:398–405.
Yang Y-S, Watson WJ, Tucker PW, Capra JD. Construction of recombinant DNA by exonuclease recession. Nucleic Acids Res. 1993;21:1889–93.
Zhao Y, Chapman DAG, Jones IM. Improving baculovirus recombination. Nucleic Acids Res. 2003;31:1–5.
Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, et al. Analysis of yeast protein kinases using protein chips. Nat Genet. 2000;26:283–9.
Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, et al. Global analysis of protein activities using proteome chips. Science. 2001;14(293):2101–5.
Zhu X, Gerstein M, Snyder M. ProCAT: a data analysis approach for protein microarrays. Genome Biol. 2006;7:110.

Jonathan Blackburn, D.Phil, Professor, South Africa Prof. Blackburn currently holds the South African Research Chair in Applied Proteomics & Chemical Biology and is head of the African Network for Drugs and Diagnostic Innovation (ANDI) Centre of Excellence in Proteomics and Genomics. He previously held a Royal Society University Research Fellowship at the Department of Biochemistry, Cambridge University, and is an EPSRC visiting research fellow at the University of Manchester.

He was founder and Research Director of the Centre for Proteomic and Genomic Research in Cape Town, was the academic founder and chief scientific officer of a UK biotechnology company, Sense Proteomic Ltd, and was chief scientist of Procognia Ltd. He obtained his DPhil degree in chemistry from the University of Oxford under the supervision of Prof. Sir Jack Baldwin, FRS, and carried out postdoctoral research at the Medical Research Council UK with Prof. Sir Alan Fersht, FRS. Prof. Blackburn serves in a number of national and international committees including the National Health Research Committee (South Africa) and the Biotechnology Subcommittee of the International Union of Pure and Applied Chemistry. He sits on the editorial advisory boards of the Journal of Proteome Research, Journal of Proteome Science and Computational Biology, and Expert Review of Proteomics. His academic expertise ranges from mechanistic enzymology, protein biochemistry, molecular biology, and proteomics to the creation of novel biomolecules by in vitro evolution. He is currently particularly interested in applications of protein microarray and mass spectrometry technologies in diagnostic marker discovery and validation, in the high-throughput study of protein-drug interactions, as well as in studying the effects of polymorphic variation on protein function.