Published in final edited form as:
Methods Mol Biol. 2006 ; 338: 245–260. doi:10.1385/1-59745-097-9:245.

Protein Binding Microarrays (PBMs) for the Rapid, High-Throughput Characterization of the Sequence Specificities of DNA Binding Proteins

Michael F. Berger[1,4] and **Martha L. Bulyk**[1,2,3,4,5]

1*Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115*

2*Department of Pathology Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115*

3*Harvard/MIT Division of Health Sciences and Technology (HST), Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115*

4*Harvard University Graduate Biophysics Program, Harvard Medical School, Boston, MA 02115*

Abstract

DNA binding proteins play a number of key roles in cells, in processes including transcriptional regulation, recombination, genome rearrangements, and DNA replication, repair, and modification. Of particular interest are the interactions between transcription factors and their DNA binding sites, as they are an integral part of the transcriptional regulatory networks that control gene expression. Despite their importance, the DNA binding specificities of most DNA binding proteins remain unknown, as earlier technologies aimed at characterizing DNA-protein interactions have been time-consuming and not highly scalable. We have developed a new DNA microarray-based technology, termed protein binding microarrays (PBMs), that allows rapid, high-throughput characterization of the *in vitro* DNA binding site sequence specificities of transcription factors in a single day. The resulting DNA binding site data can be used in a number of ways, including for the prediction of the genes regulated by a given transcription factor, annotation of transcription factor function, and functional annotation of the predicted target genes.

Keywords

DNA microarrays; transcription factors; DNA binding proteins; protein-DNA binding; DNA regulatory motifs

1. Introduction

DNA binding proteins are important in various cellular processes including transcriptional regulation, recombination, genome rearrangements, and DNA replication, repair, and modification. The interactions between transcription factors and their DNA binding sites are of particular interest because they regulate gene expression required for progression through the cell cycle, through differentiation, and in response to environmental stimuli. However, only a small handful of sequence-specific transcription factors have been characterized well enough

5Correspondence should be addressed to: Martha L. Bulyk, Ph.D., Harvard Medical School New Research Building, Room 466D, 77 Avenue Louis Pasteur, Boston, MA, 02115. Phone: (617) 525-4725. Fax: (617)525-4705. Email: E-mail: mlbulyk@receptor.med.harvard.edu..

such that all the sequences that they can, and just as importantly, can not bind, are known. This sparseness of this binding site sequence data is highly problematic because these sparse datasets are frequently used to search for genomic occurrences of these sites, with many false positive and false negative binding sites being predicted. Earlier technologies aimed at characterizing DNA-protein interactions have been time-consuming and not highly scalable, and microarray readout of chromatin immunoprecipitations (`ChIP-chip', or genome-wide location analysis) requires that the given DNA binding protein be bound to its target sites when the cells are fixed (1).

Recent advances in genomics and proteomics have set the stage for rapid, high-throughput characterization of DNA binding proteins. Overexpression and purification of DNA binding proteins of interest is a familiar technique that has been used to allow characterization of these proteins using various traditional biochemical techniques. Now, most researchers also have access to DNA microarraying facilities, if not at their own institution, then through another institution that provides microarraying services for a fee. Likewise, DNA microarray scanners and glass slides for printing of the microarrays are readily available.

We recently developed an *in vitro* DNA microarray technology, which we term protein binding microarrays (PBMs), for the characterization of the sequence specificities of DNA-protein interactions. This technology allows the *in vitro* binding specificities of individual DNA binding proteins to be determined in a single day, by assaying the sequence-specific binding of a given DNA binding protein directly to double-stranded DNA microarrays spotted with a large number of potential DNA binding sites. Specifically, a DNA binding protein of interest is expressed with an epitope tag, purified, and then bound directly to triplicate double-stranded DNA microarrays. The protein-bound microarrays are then washed to remove any nonspecifically bound protein and labeled with a fluorophore-conjugated antibody specific for the epitope tag. In order to normalize the PBM data by relative DNA concentration, separate triplicate microarrays from the same print run are stained with the dye SYBR Green I, which is specific for double-stranded DNA (*see* Figs. 1 and 2). The sequences corresponding to the significantly bound spots (*see* Fig. 3a) are analyzed with a motif prediction tool in order to identify the DNA binding site motif for the given DNA binding protein (*see* Fig. 3b). This PBM technology will likely aid in the annotation of many regulatory proteins whose DNA binding specificities have not been characterized and in the construction of gene regulatory networks. For example, binding site data derived from PBMs on transcription factors from the yeast *Saccharomyces cerevisiae*, using whole-genome *S. cerevisiae* intergenic microarrays, corresponded well with binding site specificities determined from ChIP-chip. Furthermore, comparative sequence analysis of the PBM-derived binding sites indicated that many of the sites identified as bound in PBMs, including some not identified as bound in the ChIP-chip data, are highly conserved in other *sensu stricto* yeast genomes and thus are likely to be functional *in vivo* binding sites that may be utilized in a condition-specific manner (2).

2. Materials

2.1. Preparation of Double-Stranded DNAs (dsDNAs)

1. PCR products corresponding to amplified noncoding regions of a genome of interest (*see* Note 1).
2. MultiScreen® PCR Filter Plates (Millipore, Billerica, MA) (*see* Note 2).

2.2. Printing and Processing of the dsDNA Microarrays

1. Corning® GAPS II or UltraGAPS 25 × 75 mm amino-silane coated glass slides (Fisher Scientific) (*see* Note 3).

2. Black or orange light-protective plastic slide boxes (Fisher Scientific).
3. Vacuum desiccator (Fisher Scientific) and Drierite desiccant (Fisher Scientific).
4. Microarraying facility equipped with an OmniGrid® 100 microarrayer (Genomic Solutions, Ann Arbor, MI) with Stealth 3 pins (Telechem International, Sunnyvale, CA).
5. Handheld steamer (Conair, Stamford, CT).
6. Standard lab oven (VWR International, West Chester, PA) (*see* Note 4).
7. Stratalinker® 1800 UV Crosslinker (Stratagene, La Jolla, CA).

2.3. Staining the dsDNA Microarrays

1. 2X SSC: 300 mM NaCl, 30 mM sodium citrate (*see* Note 5). Make fresh before use. Prepare using sterile water and a stock solution of 20X SSC, pH 7.0 (3), sterilized by autoclaving, and stored at room temperature.
2. SYBR Green I staining solution: 1:5000 dilution of SYBR Green I (Molecular Probes, Eugene, OR) in 2X SSC, 0.1% Triton X-100 (Fisher Scientific). Be sure that the SYBR Green I stock solution has been thawed completely at room temperature, in the dark to prevent photobleaching, and then vortexed before using. Make fresh before use.
3. Staining wash buffer: 2X SSC, 0.1% Triton X-100 (Fisher Scientific).
4. Forceps (Fisher Scientific).
5. Polypropylene Coplin staining jars (Fisher Scientific) for 3 × 1 inch slides.
6. Adjustable speed platform shaker (Fisher Scientific).
7. Table-top centrifuge and rotor suitable for centrifuging microscope slides. We use an IEC Centra CL3R equipped with a 224 Microplate Rotor (Thermo Electron, Milford, MA).

2.4. Protein Binding Microarray Experiments

1. Purified DNA binding protein, epitope-tagged with glutathione *S*-transferase (GST), stored at -80°C in PBS (*see* Note 6).

---

[1]We have recently used whole-genome yeast intergenic microarrays in PBM experiments in order to identify the DNA binding site specificities of yeast transcription factors (2). Microarrays spotted with coding regions are also expected to aid in identifying the sequence specific binding properties of DNA binding proteins, despite the fact that it is currently thought that most *in vivo* functional regulatory sites will be located in noncoding regions. Since PBM experiments are an *in vitro* technology, as long as there is sufficient sequence space represented on the DNA microarrays, one can expect to be able to derive a good approximation of the DNA binding site motif. Along these lines, one need not utilize microarrays spotted with amplicons representing genomic regions from the same genome as the DNA binding protein of interest, but rather can use microarrays spotted with a different genome's sequence. Similarly, microarrays spotted with synthetic dsDNAs can also be used in PBMs. Likewise, the dsDNAs need not be made by PCR amplification, but rather can be made by other means, such as by primer extension. We have successfully performed PBMs using microarrays spotted with PCR products whose lengths ranged from ~60 to ~1500 base pairs (2), and also using microarrays spotted with synthetic dsDNAs ranging from ~35 to ~50 base pairs (2,6).

[2]Alternatively, PCR products may be precipitated with 1 M ammonium acetate and two volumes of isopropanol, washed with 70% ethanol, dried overnight, and resuspended in 3x SSC printing buffer. The extra filtration provided by the MultiScreen® plates increases the purity of the double-stranded DNA.

[3]Any remaining, unused blank GAPS II or UltraGAPS slides from an opened package should be stored in a vacuum desiccator containing Drierite desiccant, as should all post-processed, printed microarrays.

[4]Baking the microarrays at 80°C for 2 hours in a clean oven before UV-crosslinking may improve intra-spot uniformity. However, baking may also result in a decreased shelf life of the microarrays.

[5]Unless otherwise noted, buffers are prepared using distilled, deionized water (ddH_2O).

2. Phosphate buffered saline (PBS): 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4. Adjust pH to 7.4 using 1 M HCl, sterilize by autoclaving, and then store at room temperature.

3. Zinc acetate (ZnAc) (Sigma, St. Louis, MO): 25 mM. Store in aliquots at -20°C, and then add to various buffers as required (*see* Note 7).

4. Pre-wetting buffer (PBS / 0.01% TX-100): 0.01% Triton X-100 in PBS. Make fresh before use. Prepare using a stock solution of 10% Triton X-100 (Sigma), filter-sterilized using 0.22 μm filter unit (Fisher Scientific), and stored at room temperature.

5. 2% and separately 4% nonfat dried milk (Sigma) in PBS. Allow to dissolve on a platform shaker, shaking very gently at 25 rpm, either overnight or for a few hours. Sterilize with a syringe (Fisher Scientific) equipped with a sterile 0.22 μm filter (Millipore, Billerica, MA) for the 2% milk solution, or with a 0.45 μm filter (Millipore, Billerica, MA) for the 4% milk (*see* Note 8).

6. Wash buffers: Wash buffer 1 (PBS / 0.1% Tween): 0.1% Tween-20 in PBS. Wash buffer 2 (PBS / ZnAc / 0.5% Tween): 0.5% Tween-20 in PBS containing 50 μM ZnAc. Wash buffer 3 (PBS / ZnAc / 0.05% Tween): 0.05% Tween-20 in PBS containing 50 μM ZnAc. Wash buffer 4 (PBS / ZnAc / 0.01% TX-100): 0.01% Triton X-100 in PBS containing 50 μM ZnAc. Make all four wash buffers fresh before use. Prepare using a stock solution of 20% Tween-20 (Sigma) or 10% Triton X-100 (Sigma), filter-sterilized using 0.22 μm filter unit (Fisher Scientific), and stored at room temperature.

7. Salmon testes DNA (Sigma).

8. Bovine serum albumin (New England Biolabs, Beverly, MA).

9. Alexa Fluor® 488 conjugated anti-glutathione *S*-transferase (anti-GST) polyclonal antibody (Molecular Probes) (*see* Note 6).

10. LifterSlips™ cover slips (Erie Scientific, Portsmouth, NH).

11. Kimwipes (Fisher Scientific).

12. Hydration chamber (*see* Note 9).

13. Polypropylene Coplin staining jars (Fisher Scientific) for 3 × 1 inch slides.

14. Adjustable speed platform shaker (Fisher Scientific).

15. Table-top centrifuge and rotor suitable for centrifuging microscope slides. We use an IEC Centra CL3R equipped with a 224 Microplate Rotor (Thermo Electron, Milford, MA).

16. Diamond-tipped glass scribe (Fisher Scientific).

---

[6]Purified DNA binding proteins should be aliquoted before storing at -80°C, in order to avoid unnecessary freeze/thaw. Antibody should be stored according to manufacturer's recommendations. For long-term storage of Alexa Fluor® 488 conjugated anti-glutathione *S*-transferase (anti-GST) polyclonal antibody (Molecular Probes), we recommend aliquoting and storing at -20ºC, per manufacturer's recommendations. However, we have observed no noticeable decrease in signal intensity upon storing this antibody at 4ºC for one year.

[7]ZnAc is necessary only when performing PBM experiments on zinc finger proteins.

[8]Filtering of the 2% and 4% milk solutions serves two purposes: (1) sterilization; and (2) removal of fine particulates that may contribute to noise in the PBM data. We have found that the 4% milk solution readily clogs a 0.22 μm filter unit, and so we recommend using a 0.45 μm filter unit for syringe-filtering of the 4% milk solution. The 2% milk solution can be syringe-filtered with a 0.22 μm filter. Alternatively, 0.45 μm filters can be used for sterilizing both the 2% and 4% milk.

[9]Microarrays are incubated in a hydration chamber to prevent excessive evaporation of the reaction mixture under the cover slip. An empty pipet tip box works nicely. Lift out the tip rack, fill the bottom of the pipet tip box with about half an inch of sterile water, and replace the tip rack. Wipe off the inside of the lid and the tip rack with ethanol using a Kimwipe before every use, and between reaction steps in the PBM experiments.

2.5. Scanning of the Microarrays

1. ScanArray 5000 microarray scanner equipped with argon ion laser (488 nm excitation) and 522 nm emission filter (Perkin Elmer, Boston, MA).
2. Computer running Windows NT or Windows 2000 and equipped with at least 128 MB RAM, to run the above microarray scanner using ScanArray microarray analysis software.
3. Kimwipes (Fisher Scientific).
4. Canned air (Fisher Scientific).

2.6. Quantification of the Microarray Signal Intensities

1. GenePix microarray analysis software installed on an IBM-compatible computer with a 1 GHz Pentium or faster, Windows 2000 or XP operating system, 512 MB RAM, and at least 12 MB disk space.
2. Masliner software. Masliner is a Perl script, so it can be run either locally under DOS box command lines on Win32 systems, or on a Linux cluster. (http://arep.med.harvard.edu/masliner/pgmlicense.html.)
3. Excel software (Microsoft).

2.7. Post-Quantification Data Analysis

1. R statistics package (www.r-project.org) installed on a Linux, Macintosh, Unix, or Windows computer. The minimal system requirements for Windows are: Windows 95/98/ME/NT4/2000/XP Server, 16 MB RAM, and 50 MB disk space.
2. Mathematica software package (Wolfram Research, Inc., Champaign, IL) installed on a Windows, Macintosh, or Linux computer. For Windows, the minimal system requirements are: 128 MB RAM and 500 MB disk space, and it is compatible with Windows 98/Me/NT 4.0/2000/XP.
3. BioProspector software (4) (http://motif.stanford.edu/distributions/). BioProspector is a C program. On our Linux machines, BioProspector takes up 113 KB.
4. ScanACE and MotifStats software (5) (http://atlas.med.harvard.edu/download/extra.html). These are both C++ programs. On our Linux machines, ScanACE takes up 1.6 MB and MotifStats takes up 115 KB.

3. Methods

2.1. Preparation of Double-Stranded DNAs (dsDNAs)

1. Amplify genomic regions by PCR. For the whole-genome yeast intergenic microarrays (2), we performed 25 cycles at 94°C for 30 seconds, 60°C for 30 seconds, and 72°C for 90 seconds.
2. Filter PCR products using a 96-well MultiScreen® PCR filter plate according to manufacturer's protocols. After application of a vacuum for 10 minutes, plates may be air-dried in a clean chemical hood (*see* Note 2.)

2.2. Printing and Processing of the dsDNA Microarrays

1. Suspend dsDNA in approximately 15 μl 3x SSC spotting buffer at a concentration between 100 and 500 ng/μl. Spot DNA onto GAPS II or UltraGAPS slides using an OmniGrid® 100 microarrayer. Approximately 0.7 nl will be deposited at each spot.

2. Lay slides face-up and flat, and blow a moisture stream across their surface for approximately 5 seconds using a handheld steam humidifier. Alternatively, one can incubate the microarrays in a humid chamber at 37°C for 48 hours.
3. Bake microarrays in a standard VWR lab oven at 80°C for 2 hours (*see* Note 4).
4. Apply 300 mJ to each slide using a Stratalinker® UV Crosslinker in order to crosslink the dsDNA to the substrate surface.

2.3. Staining the dsDNA Microarrays

1. Prepare SYBR Green I staining solution (*see* Note 10), and mix before using to stain microarrays.
2. Stain the microarrays using SYBR Green I staining solution for 12 min by shaking in a Coplin jar at room temperature @ ~100-125 rpm on a platform shaker.
3. Wash the microarrays in 2X SSC, 0.1% Triton X-100 staining wash buffer for 5 min at room temperature (*see* Note 11).
4. Wash the microarrays in 2X SSC for 2 minutes at room temperature.
5. Immediately spin slides dry in a table-top centrifuge by centrifuging for 5 min at 40 x *g* (500 rpm using an IEC Centra CL3R).

2.4. Protein Binding Microarray Experiments

1. Using a glass scribe, etch small grooves on the face (i.e., DNA side) of each microarray, at a distance of a few millimeters beyond the borders of the printed area. This will help to confine all solutions to the center portion of the microarray throughout the protein binding microarray experiment.
2. Pre-wet the microarrays in PBS / 0.01% TX-100 for at least 5 min by shaking in a Coplin jar at room temperature @ ~125 rpm on a platform shaker.
3. While the microarrays are being pre-wet, thaw the previously purified DNA binding protein of interest on ice.
4. Working with one microarray at a time, quickly remove the microarray from the Coplin jar, gently shake off any excess buffer, and wipe the back (i.e., the non-DNA side) and sides of the microarray with a Kimwipe (*see* Note 12).
5. Apply (*see* Notes 13 and [14]) 250 μl of 2% milk solution (pre-blocking buffer) to the microarrays, cover with a LifterSlips™ cover slip, and allow to incubate in a hydration chamber for 1 hour at room temperature.

---

[10]Note that SYBR Green I and Alexa Fluor® 488 conjugated anti-glutathione *S*-transferase (anti-GST) polyclonal antibody are light-sensitive, and so measures should be taken to avoid their photobleaching in the course of the SYBR Green I straining and PBM experiments. We recommend turning off all overhead and benchtop lighting when handling these reagents, and also when handling the microarrays once they have been stained with either SYBR Green I or the fluorophore-conjugated antibody.

[11]All wash steps in Coplin jars are performed by shaking at room temperature @ ~125 rpm on a platform shaker. Shaking at speeds faster than ~125 rpm may cause the Coplin jar to tip over while shaking.

[12]Drying the back and sides of the microarray with a Kimwipe helps to prevent solution from leaking out from the edges of the cover slip. Because of the hydrophobicity of the active surface of the Corning® GAPS II and UltraGAPS glass slides, drying the front of the glass slide outside the perimeter of the DNA spots, using the etched grooves as a guide, can help to confine the protein or antibody solution, once dispensed onto the microarray, to the area of the microarray that contains the DNA spots. This must be done quickly so that the area containing the DNA spots does not dry.

[13]Briefly centrifuge all reaction mixtures before applying to microarrays in order to remove bubbles. When pipeting the reaction mixtures onto the microarrays, avoid pipeting up any fine bubbles that may remain at the very top surface of the reaction mixtures. If nevertheless a few air bubbles become apparent once a reaction mixture has been applied to a microarray, very carefully attempt to pipet up the bubbles, while avoiding the removal of the reaction mixture. If some air bubbles still remain, they may be brought to the edge of the glass slide, and thus outside the spotted area, by gently rocking the cover slip as it is laid down on the microarray.

6. As soon as the microarray pre-blocking has been set up, pre-incubate the DNA binding protein of interest with nonspecific competitors, for 1 hour at room temperature. Specifically, dilute the thawed DNA binding protein to a 20 nM final concentration in a 100 μl protein binding reaction mixture consisting of PBS, 50 μM ZnAc, 2% (w/v) nonfat dried milk (Sigma), 51.3 ng/μl salmon testes DNA (Sigma), and 0.2 μg/μl BSA. The final 2% milk concentration is achieved by using a 2-fold dilution of the 4% milk solution that was prepared. A 100 μl reaction volume will be adequate for a printed microarray that encompasses ~2/3 of the slide surface.

7. While the microarrays and the DNA binding protein are pre-blocking separately, thaw the Alexa Fluor® 488 conjugated anti-GST polyclonal antibody on ice, covered with either an ice bucket lid or aluminum foil, in order to prevent photobleaching (*see* Note 10).

8. Once the 1 hour pre-blocking step is completed, wash the microarrays once with PBS / 0.1% Tween for 5 min (*see* Note 8), followed by once with PBS / ZnAc / 0.01 % TX100 for 2 min.

9. Wipe the back and sides of the microarrays with a Kimwipe, apply the protein binding reaction mixture to the microarrays, cover with a LifterSlips™ cover slip, and allow to incubate in a hydration chamber for 1 hour at room temperature.

10. As soon as the microarray protein binding reaction has been set up, dilute the Alexa Fluor® 488 conjugated anti-GST polyclonal antibody to a concentration of 0.05 mg/ml in 2% milk in 150 μl of PBS containing 50 μM ZnAc, and allow to pre-incubate for 1 hour at room temperature in the dark.

11. Once the 1 hour protein binding step is completed, wash the microarrays once in PBS / ZnAc / 0.5% Tween for 3 min, followed by once in PBS / ZnAc / 0.01% TX100 for 2 min.

12. Wipe the back and sides of the microarrays with a Kimwipe, apply the pre-incubated antibody mixture to the microarrays, cover with a LifterSlips™ cover slip, and allow to incubate in a hydration chamber for 1 hour at room temperature, covered with either an ice bucket or aluminum foil, in order to protect from light.

13. Once the 1 hour antibody staining is completed, wash the microarrays three times with PBS / ZnAc / 0.05 % Tween, with each wash going for 3 min, followed by once with PBS / ZnAc for 2 min.

14. Immediately spin slides dry in a table-top centrifuge by centrifuging for 6 min at 40 x *g* (500 rpm using an IEC Centra CL3R).

2.5. Scanning of the Microarrays

1. Wipe the backs (i.e., the non-DNA sides) of the microarrays with a slightly dampened Kimwipe, in order to remove any streaks or spots due to dried buffer.

2. Very quickly blow any lint off of the spun-dry microarrays using canned air.

3. Scan the microarrays at a range of different laser power intensities or photomultiplier tube (PMT) gain settings per microarray, using an appropriate laser and filter set (488 nm excitation / 522 nm emission for SYBR Green I and Alexa Fluor™ 488). Typically

[14]In applying reaction mixtures onto the microarrays, certain techniques can aid in spreading the mixture over the surface of the microarray, and in increasing the homogeneity of the reaction mixture when applied to a pre-wet or washed microarray. The reaction mixture can be dispensed one droplet at a time, covering the entire surface where the DNA was spotted. The microarray can also be rocked back and forth to spread the reaction mixture uniformly across the spotted area. The use of LifterSlips™ cover slips helps to assure a uniform distribution of the reaction mixture over the surface of the microarray.

~3-6 different settings are used (*see* Note 15), so that signal intensities are captured for even very low signal intensity spots, while ensuring that we captured sub-saturation signal intensities for each of the spots on the microarray (2,6).

2.6. Quantification of the Microarray Signal Intensities

1. Quantify microarray TIF images with GenePix Pro software (Axon Instruments, Inc.). Analyze as a single-color image, and keep the feature size fixed throughout the alignment procedure.
2. Using Excel, GenePix, or other software, calculate the background-subtracted median intensities using the median local background.
3. Use masliner (MicroArray Spot LINEar Regression) software (7) to calculate the relative signal intensities over the full series of laser power (or PMT gain) setting scans in a semiautomated fashion. Masliner combines the linear ranges of multiple scans from different scanner sensitivity settings onto an extended linear scale (2,6, 7). The dynamic range of the final PBM and SYBR Green I stained microarrays frequently have post-masliner fluorescence intensities that span 5 to 6 orders of magnitude (2).

2.7. Post-Quantification Data Analysis

2.7.1 Microarray Data Quality Control (see Note 16 and Fig. 4)—

1. For microarray data on a given DNA binding protein, for each of the triplicate protein binding microarrays, remove data corresponding to any flagged spots (i.e., spots that had dust flecks, etc.).
2. Normalize the data from each of three triplicate microarrays according to total signal intensity, so that the average spot intensity is the same for all three microarrays.
3. Within each individual microarray, separate the data into sectors, according to their local region on the slide. For example, for the whole-genome yeast intergenic arrays (2), we sectored the spots into the 32 subgrids of the printed microarray.
4. Normalize the data again so that the mean spot intensity is the same over all the sectors. This serves to normalize for any region-specific inhomogeneities in the background and also binding and labeling reactions.
5. Remove any spots whose standard deviation (SD) divided by median value is greater than 2, i.e., spots with highly variable pixel signal intensities.
6. Average the background-subtracted, normalized signal intensities for all spots with reliable data in at least two of the three replicate microarrays, and calculate the SD/mean value. Remove any spots for which the SD/mean value is greater than 1.
7. Treat the SYBR Green I microarray data exactly the same way, except here remove any spots with fewer than 50% pixels with signal intensities greater than two SDs beyond the median background signal intensity, as these spots presumably do not

[15]For our ScanArray 5000 microarray scanner, we have found the PMT gain to be optimal between 70 and 80%. We typically fix the PMT gain setting and vary the laser power in increments of 10-15% (in terms of total laser power) such that there are no spots with saturated signal intensities in the lowest intensity scan.

[16]It is important to ensure that each spot has enough DNA present to allow the accurate quantification of its signal intensity, which is consequently used to estimate the degree of sequence-specific binding of a given DNA binding protein to that spot. If the DNA concentration at a particular spot is too low or if the DNA is spread non-uniformly throughout the pixels of a particular spot, accurate measurements are more difficult. For this reason, it is important to remove such error-prone spots from consideration. Since some spots may be noisy (i.e., spots with highly variable pixel signal intensities) even after the use of this filter, we also remove noisy spots from consideration.

have enough DNA present to allow accurate quantification of signal intensities (2) (*see* Note 17).

2.7.2 Identification of the `Bound' Spots—

1. Calculate the $\log_2$ ratio of the mean PBM signal intensity divided by the mean SYBR Green I signal intensity, and create a scatter plot of the log ratio versus the spots' SYBR Green I signal intensities.
2. Although we expect that the log ratio should be independent of DNA concentration, we have found that higher DNA concentrations, as determined by higher SYBR Green I signal intensities, appear to bind proportionately less protein. In order to restore the independence of log ratio and SYBR Green I intensity, fit the scatter plot with a locally weighted least squares regression using the LOWESS function (smoothing parameter = 0.5) (8) of the R statistics package.
3. Subtract the value of the regression at each spot from its log ratio, yielding a modified log ratio that is independent of DNA concentration.
4. Plot the distribution of all log ratios as a histogram (bin size = 0.05), which is expected to resemble a Gaussian distribution with a heavy tail.
5. Determine the mode of the distribution by searching for the window of nine bins with the highest number of spots and taking the middle bin.
6. Reflect all values less than the mode and fit these values to a Gaussian function using the Mathematica software package. This provides the mean and SD of the distribution of nonspecifically-bound spots.
7. Adjust the log ratios so that the peak of the distribution of nonspecifically-bound spots is centered on zero.
8. Calculate a *P*-value for each spot based on *z*, the number of standard deviations that the spot's log ratio departs from the mean of the Gaussian distribution, using the normal error integral (9) (*see* Note 18). The *P*-value can be calculated easily in Microsoft Excel using the standard normal cumulative distribution function: normsdist(-*z*). This *P*-value for each spot represents the probability that the spot is contained within the distribution of nonspecifically-bound spots. Thus, spots with very small *P*-values in the heavy upper tail of the real distribution are likely to be bound sequence-specifically by the given DNA binding protein.
9. In order to correct for multiple hypothesis testing, adjust all individual *P*-values to a modified significance level using the Modified Bonferroni Method (10,11). For significance testing of the PBM data, we recommend using an initial $\alpha = 0.001$, which corresponds to α' equal to approximately 1.5×10^{-7} for the highest-ranking test case when evaluating ~6400 unique spots, which is the case for typical yeast intergenic microarray data (2). Spots meeting or exceeding α' are considered `bound' at a statistically significant threshold (*see* Fig. 3a).

[17]We found empirically that the following three additional filtering criteria helped to eliminate `false positive' calls (i.e., spots with no identifiable binding sites being erroneously identified as bound): (1) DNA length greater than 1500 bp; (2) low SYBR Green I raw signal intensity; and (3) low DNA density (SYBR Green I / length). These three additional filters together removed 2.7% of spots from consideration in our PBM experiments using yeast whole-genome intergenic microarrays (2). Here we are not providing the actual values for the second and third filtering criteria because these values will vary somewhat among individual microarray scanners. We recommend that the user use all three of these criteria as suggested guidelines to employ and adjust as may be appropriate.

[18]This function is related to the probability of observing a data point greater than *z* standard deviations above the mean of a normal distribution. Strictly speaking, we are not calculating a true *z*-score, since here we do not calculate the *P*-value relative to all the data, but rather just to the reflected left-half of the distribution.

2.7.3 Discovery of the DNA Binding Site Motif—

1. To search for motifs that are over-represented in PBM experiments and thus are the likely DNA binding site motifs of the given DNA binding protein, select the sequences from all the spots that had a Bonferroni-corrected *P*-value less than or equal to 0.001.
2. For this set of input sequences, use BioProspector (4) (*see* Note 19) to perform separate motif searches at each width between 6 and 18 nucleotides in order to identify the highest scoring motifs at each width.
3. Identify all matches to the motif within all sequences spotted on the microarray, and then calculate the group specificity score (5) of each discovered motif. These tasks can be accomplished with the pair of programs ScanACE and MotifStats (5), or with the software package MultiFinder (Huber and Bulyk, manuscript in preparation).
4. Choose the single motif with the lowest group specificity score (5) to be the most significant, using the set of all sequences spotted on the microarray as the background. We use this scoring metric because it indicates the degree to which the property of containing the sequence motif is specific to the input set of intergenic regions, as determined from the most significantly bound spots on the microarrays. A lower, and thus better, group specificity score indicates that the motif is more specific to the input set of spots (i.e., the spots beyond a 0.001 *P*-value threshold in the PBM data, or the randomly selected spots in the computational random controls (see below)).
5. In order to assess the statistical significance of the DNA sequence motifs resulting from analysis of the PBM experiments, perform a set of computational negative control motif searches. Specifically, perform identical motif searches on 10 separate sets of randomly selected spots from the same microarrays used to perform the PBM experiments, with each of the 10 random sets containing the same number of sequences as the original input set for the given PBM dataset.
6. Motifs with group specificity scores that are more significant as compared to the group specificity scores of the corresponding computational negative control sets are considered to indicate that the given motif is likely to correspond to the DNA binding site motif for the given DNA binding protein (*see* Fig. 3b). Examples of the ranges of group specificity scores for computational negative controls and for actual PBM data for yeast transcription factors can be found in ref. 2.

Acknowledgements

We thank Tom Volkert for technical assistance. This work was supported in part by National Institutes of Health grants from the National Human Genome Research Institute to M.L.B. (R01 HG002966 and R01 HG003420). M.F.B. was supported in part by a Graduate Research Fellowship from the National Science Foundation.

References

1. Bulyk M. Computational prediction of transcription-factor binding site locations. Genome Biol 2003;5:201. [PubMed: 14709165]
2. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet 2004;36:1331–1339. [PubMed: 15543148]
3. Sambrook, J.; Fritsch, E.; Maniatis, T. Molecular Cloning: A Laboratory Manual. Vol. 2nd Ed.. Vol. 3 vols.. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 1989.

[19]Other motif finders, such as AlignACE (5,12), MEME (13), and MDscan (14), can also be used to identify the DNA binding site motif. We chose BioProspector over other available motif finding programs because it proved to be the most inclusive in accepting the largest number of input sequences in construction of yeast transcription factor binding site motifs (2).

4. Liu X, Brutlag D, Liu J. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput 2001:127–138. [PubMed: 11262934]

5. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J. Mol. Biol 2000;296:1205–1214. [PubMed: 10698627]

6. Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc. Natl. Acad. Sci. USA 2001;98:7158–7163. [PubMed: 11404456]

7. Dudley A, Aach J, Steffen M, Church G. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. U.S.A 2002;99:7554–7559. [PubMed: 12032321]

8. Cleveland W, Devlin S. Locally weighted regression: An approach to regression analysis by local fitting. J. American Statistical Association 1988;83:596–610.

9. Taylor, J. An Introduction to Error Analysis. Vol. 2nd Ed.. University Science Books; Sausalito, CA: 1997.

10. Bulyk M, Johnson P, Church G. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 2002;30:1255–1261. [PubMed: 11861919]

11. Sokal, R.; Rohlf, R. Biometry: The Principles and Practice of Statistics in Biological Research. Vol. Third Ed.. W. H. Freeman and Company; New York: 1995.

12. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol 1998;16:939–945. [PubMed: 9788350]

13. Bailey T, Elkan C. The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol 1995;3:21–29. [PubMed: 7584439]

14. Liu X, Brutlag D, Liu J. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol 2002;20:835–839. [PubMed: 12101404]

15. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990;18:6097–6100. [PubMed: 2172928]

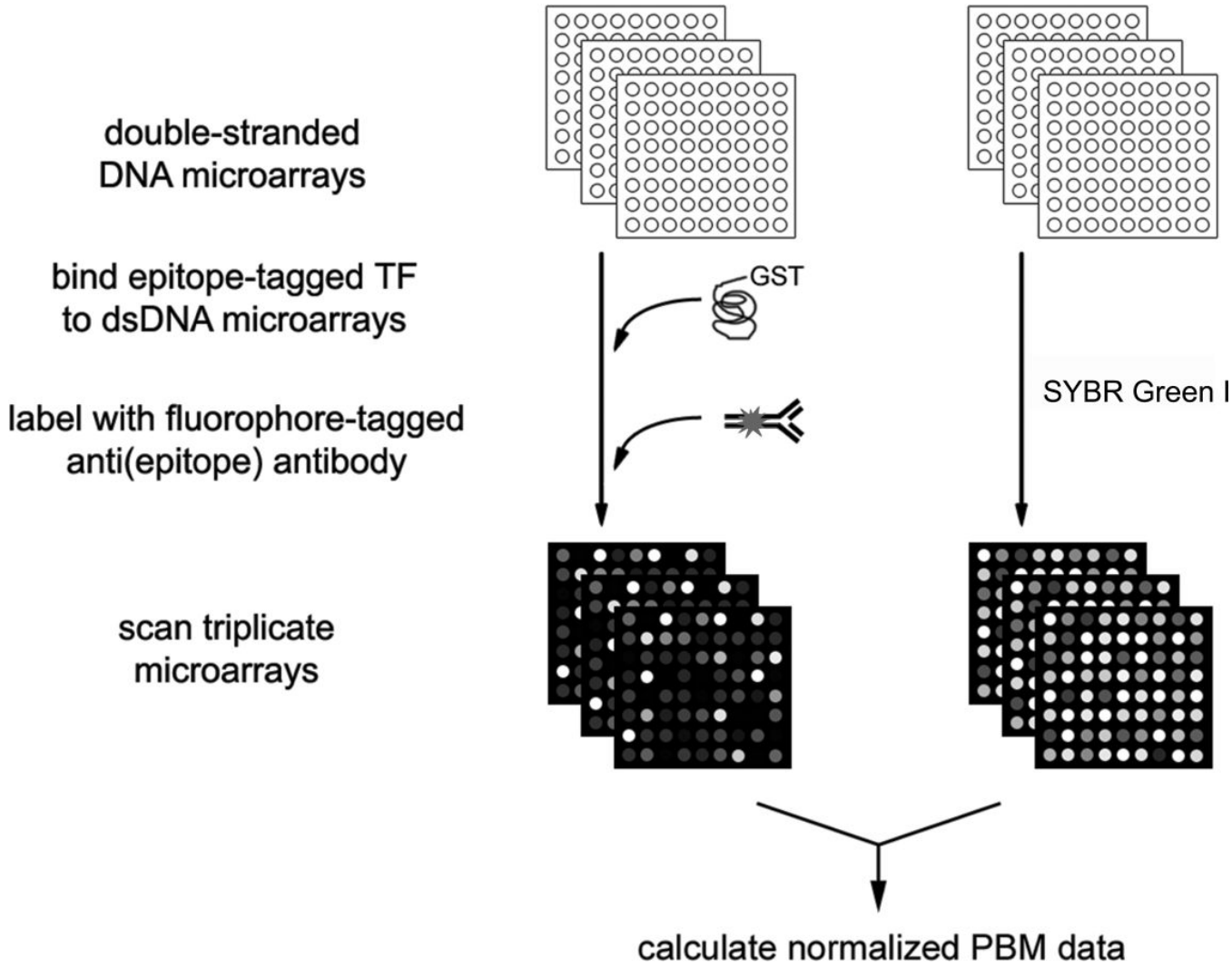

Figure 1.
Schema of protein binding microarray experiments. (Reproduced from ref. 2 with permission from Nature Publishing Group.)

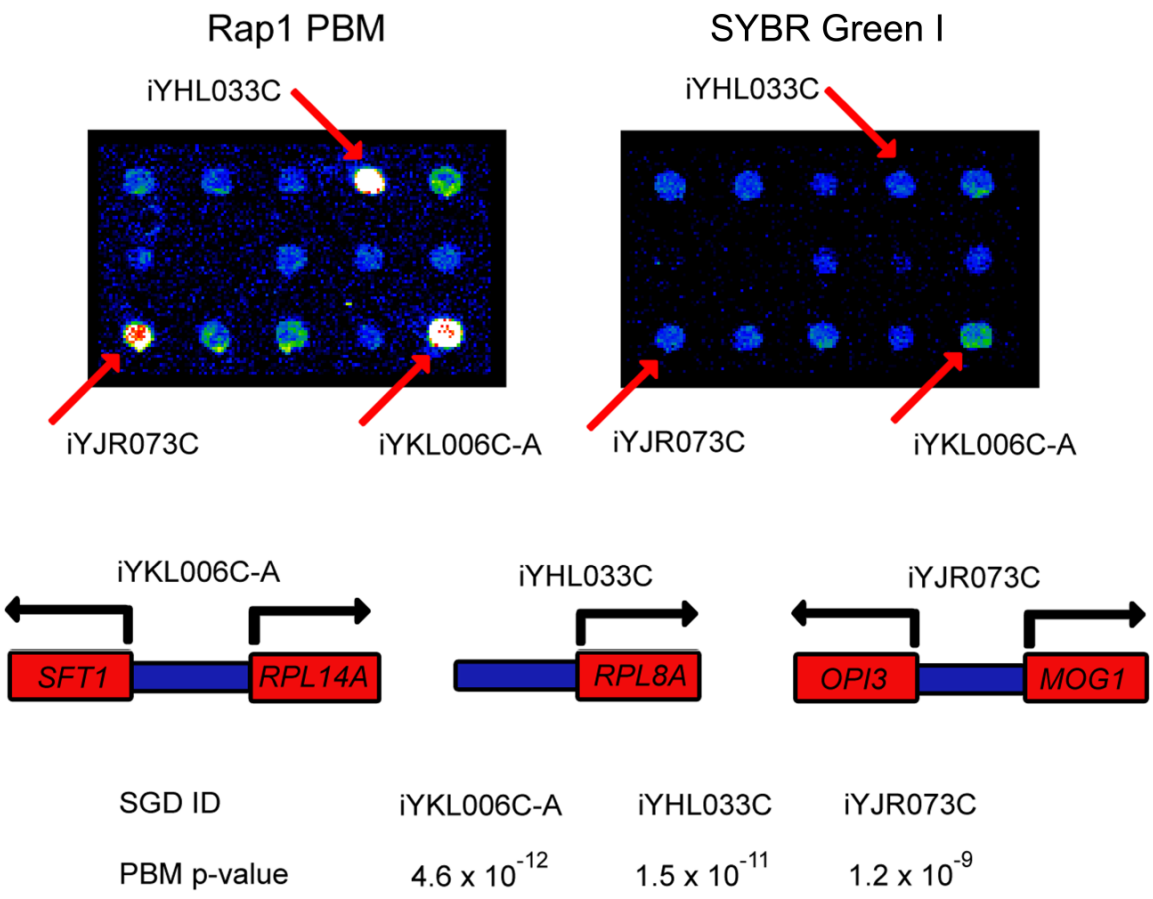

Figure 2.

Magnification of identical portions of a yeast intergenic microarrays used in a PBM experiment (left panel) or stained with SYBR Green I (right panel). Fluorescence intensities are shown in false color, with white indicating saturated signal intensity, red indicating high signal intensity, yellow and green indicating moderate signal intensity, and blue indicating low signal intensity. The three labeled spots correspond to the intergenic regions depicted below, along with the *P*-values derived from triplicate PBM and SYBR Green I microarray data. (Reproduced from ref. 2 with permission from Nature Publishing Group.)

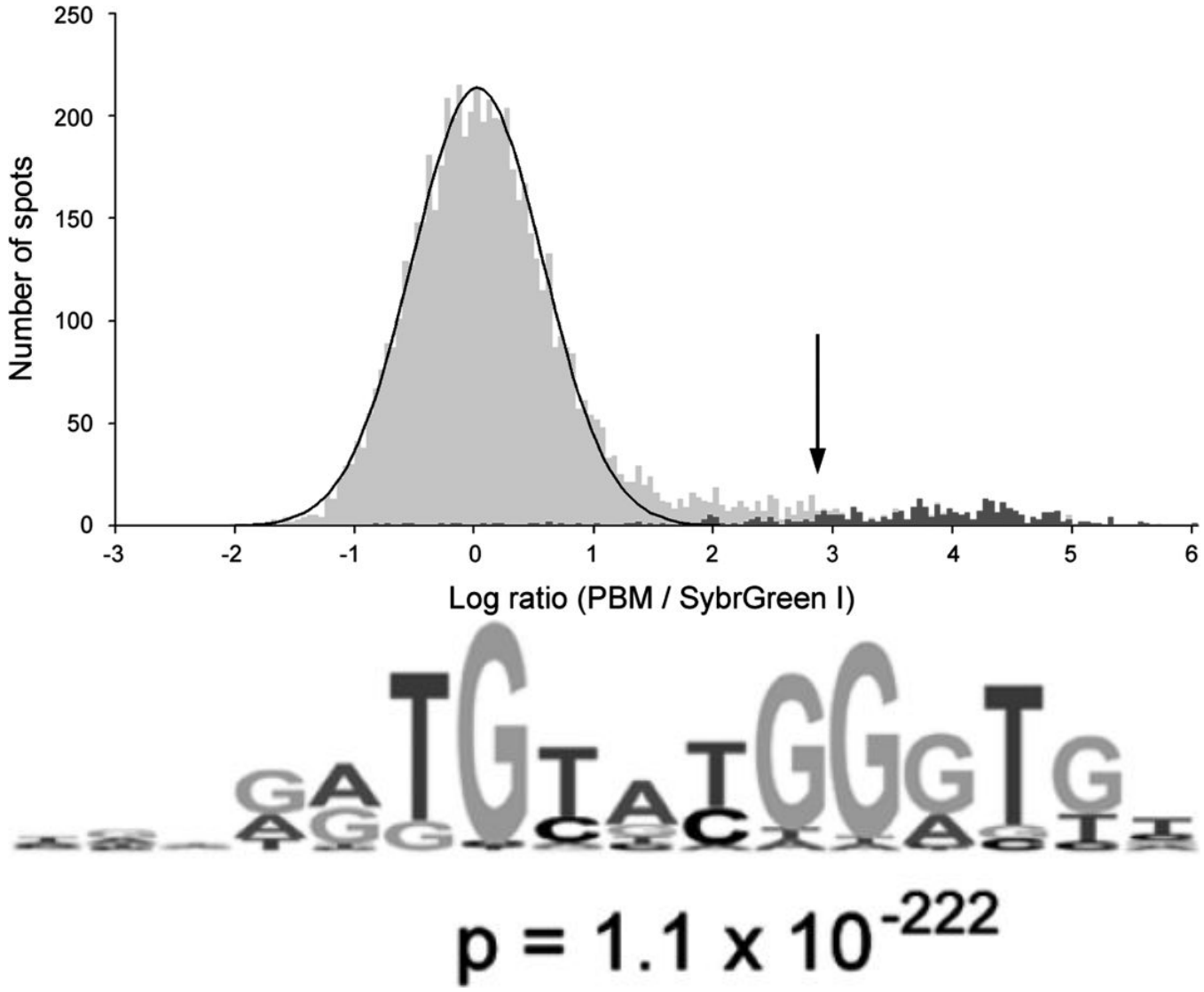

Figure 3.

Identification of the DNA binding site motif from the significantly bound spots. **(a)** Distribution of ratios of PBM data, normalized by SYBR Green I data, for the yeast transcription factor Rap1 bound to yeast intergenic microarrays. The arrow indicates those spots passing a *P*-value cutoff of 0.001 after correction for multiple hypothesis testing. Indicated in dark gray are spots with an exact match to a sequence belonging to the PBM-derived binding site motif. **(b)** Sequence logo (15) of the PBM-derived motif for the yeast transcription factor Rap1. (Reproduced from ref. 2 with permission from Nature Publishing Group.)

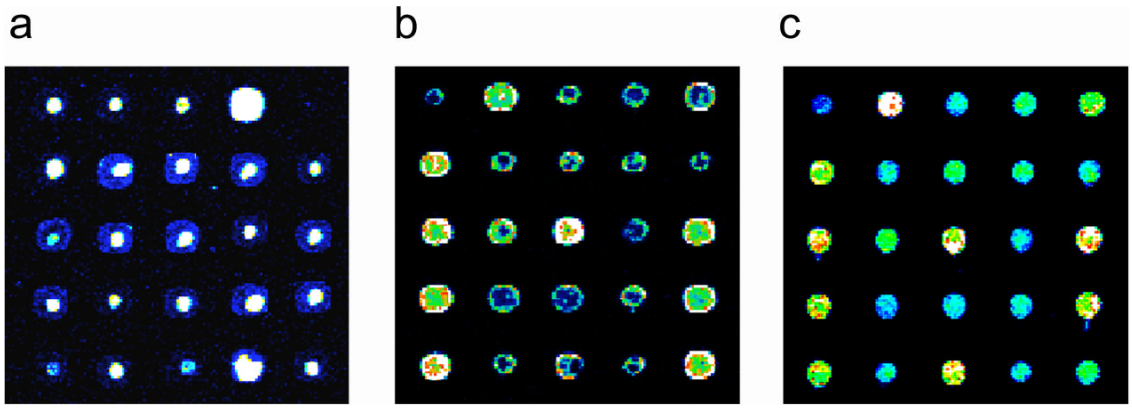

Figure 4.

Examples of DNA microarray spot quality. Identical portions of yeast intergenic microarrays printed onto Corning® GAPS II slides, processed in different ways (see below) before UV crosslinking, and then stained with SYBR Green I. Images have been false-colored as in Figure 2. Examples of microarrays with poor spot quality are shown in **(a)** and **(b)**. In both of these cases, the DNA is distributed non-uniformly, with either **(a)** high concentrations near the centers of spots, or **(b)** high concentrations along spot perimeters. Both of these microarrays resulted from two separate print runs, from which microarrays were UV crosslinked without first rehydrating and baking. An example of a good quality microarray is shown in **(c)**. This microarray was rehydrated and then baked before being UV crosslinked.