GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
`kabagg@mdanderson.org`
`kcoombes@mdanderson.org`

1 November 2005

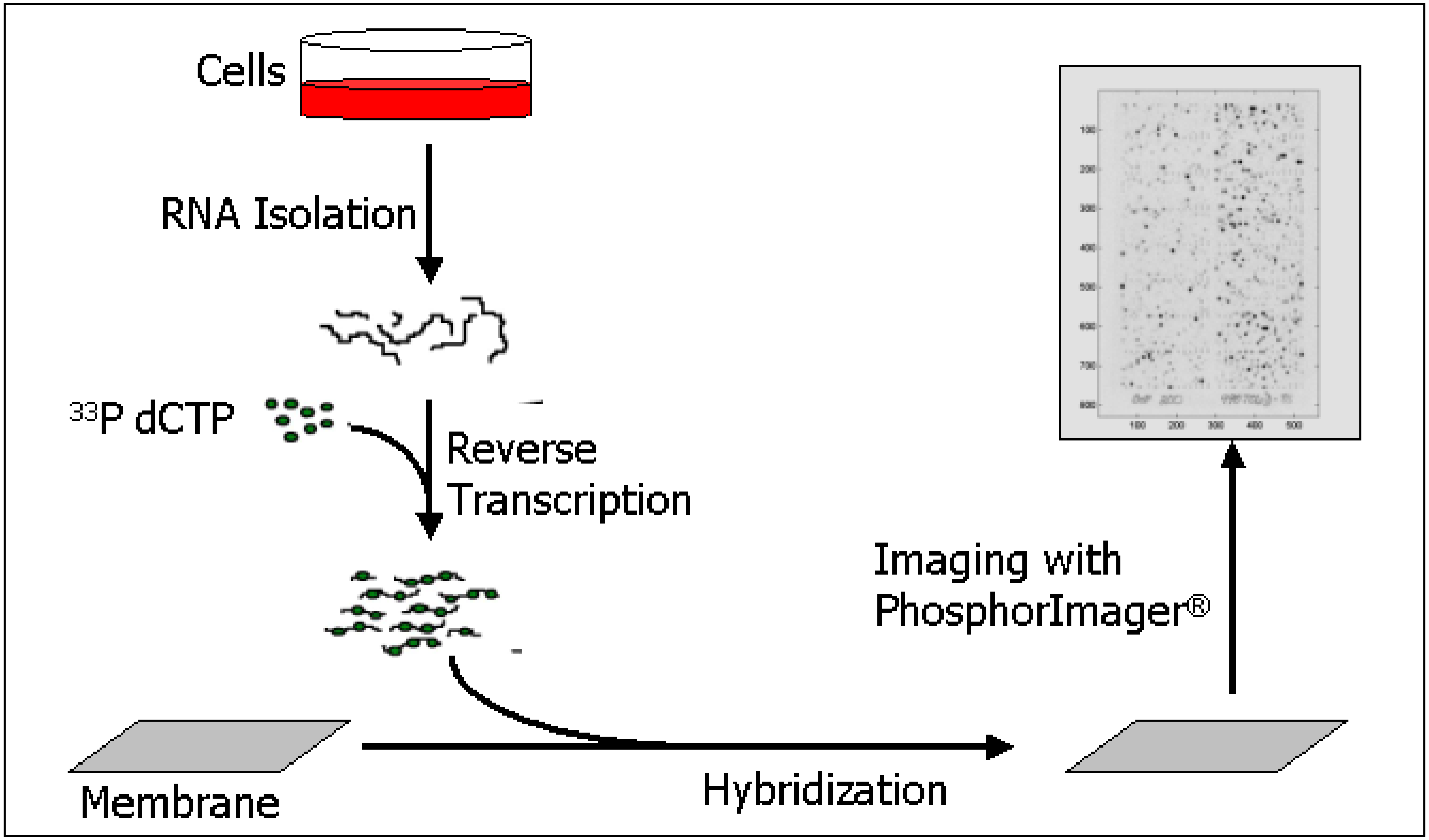
Lecture 17: Quantifying Glass Microarrays

- Microarray Platforms
- Overview: Two-color Spotted Microarrays
- TIFF files: the good and the bad
- Counting pixels: foreground, shapes, and masks
- The nothing that is: background
- Summarizing the spot: log ratios
- The effects of preprocessing: background subtraction

Microarray Platforms

- Multiple synthesized short oligos (25-mers) on silicon
  - Commerically produced: Affymetrix
  - Single channel fluorescent labeling
  - Between 11 and 20 probes per gene target
- Spotted cDNA on nylon membranes (obsolete)
  - Commerically produced: Research Genetics, Clontech
  - Radioactive labeling, single channel
- Spotted cDNA or long oligos (60- or 70-mers) on glass slides
  - Home-grown or commercial
  - Two-channel: simultaneous co-hybridization of two samples
  - Two-color fluorescent labeling

Overview: Nylon cDNA Microarrays



Cells

RNA Isolation

^{33}P dCTP

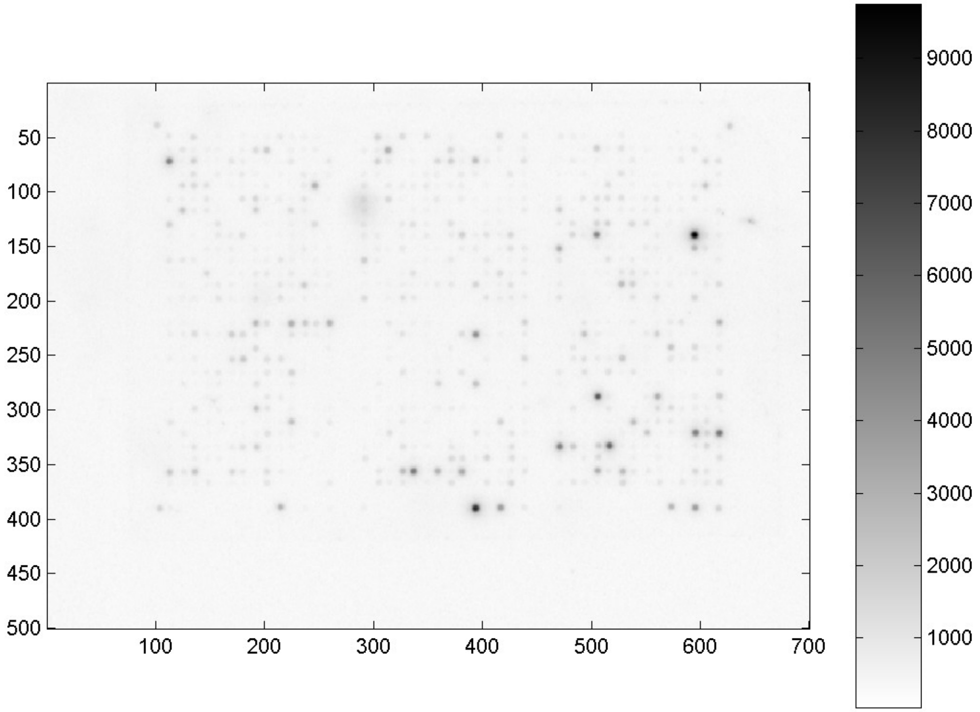
Reverse Transcription

Imaging with PhosphorImager®

Membrane

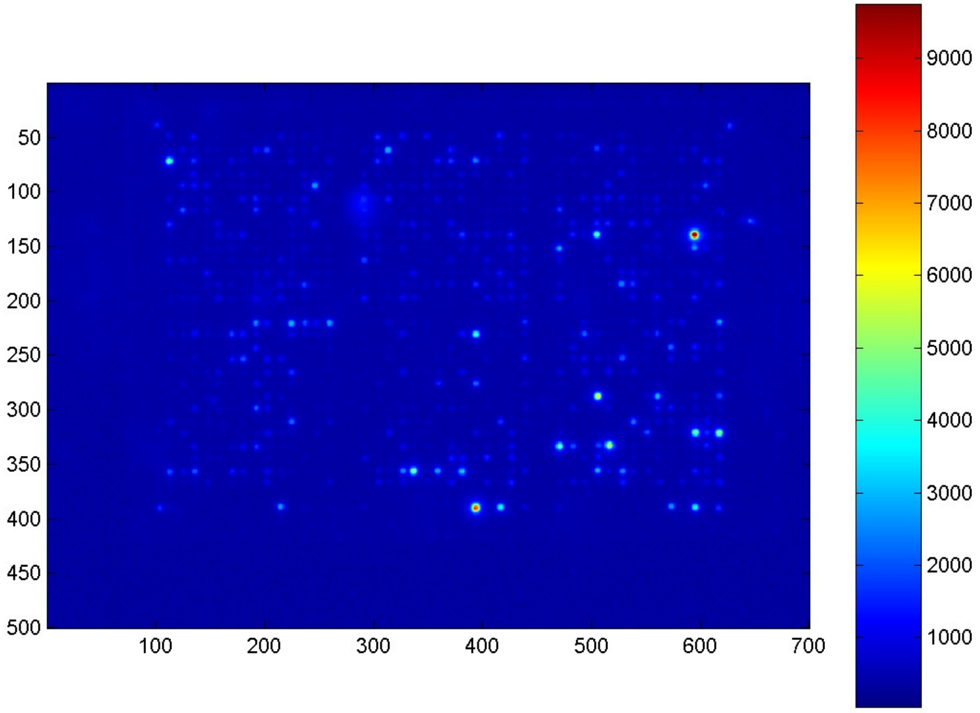
Hybridization

Nylon cDNA Microarrays



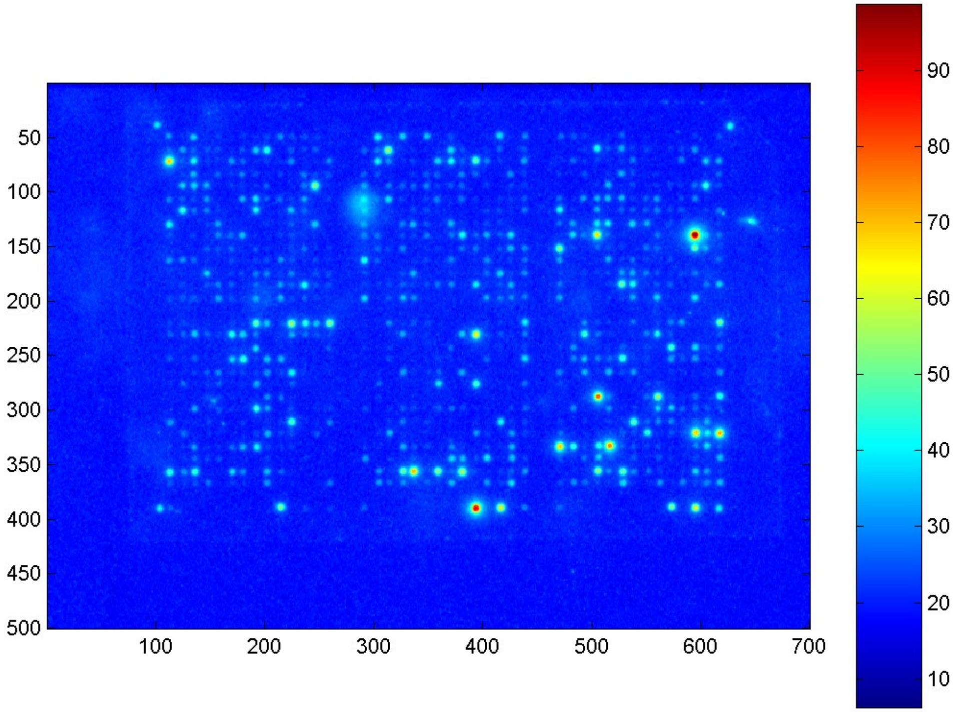
The raw data is a 16-bit, gray-scale, TIFF image.

Nylon cDNA Microarrays



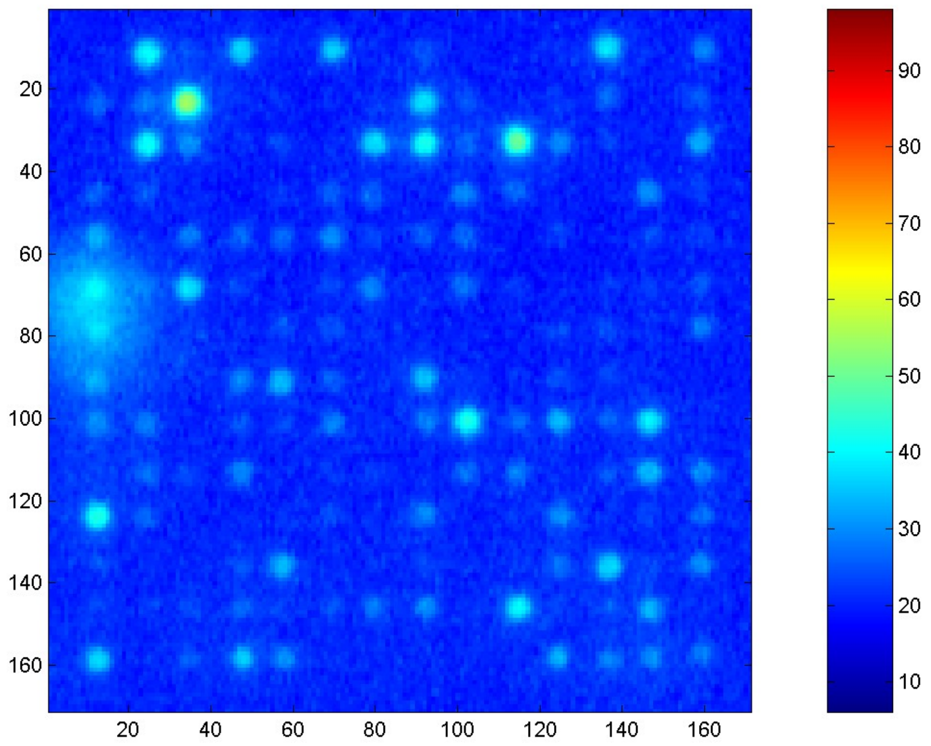
Array images are often viewed in color by changing the color map; this does not change the actual data.

Nylon cDNA Microarrays



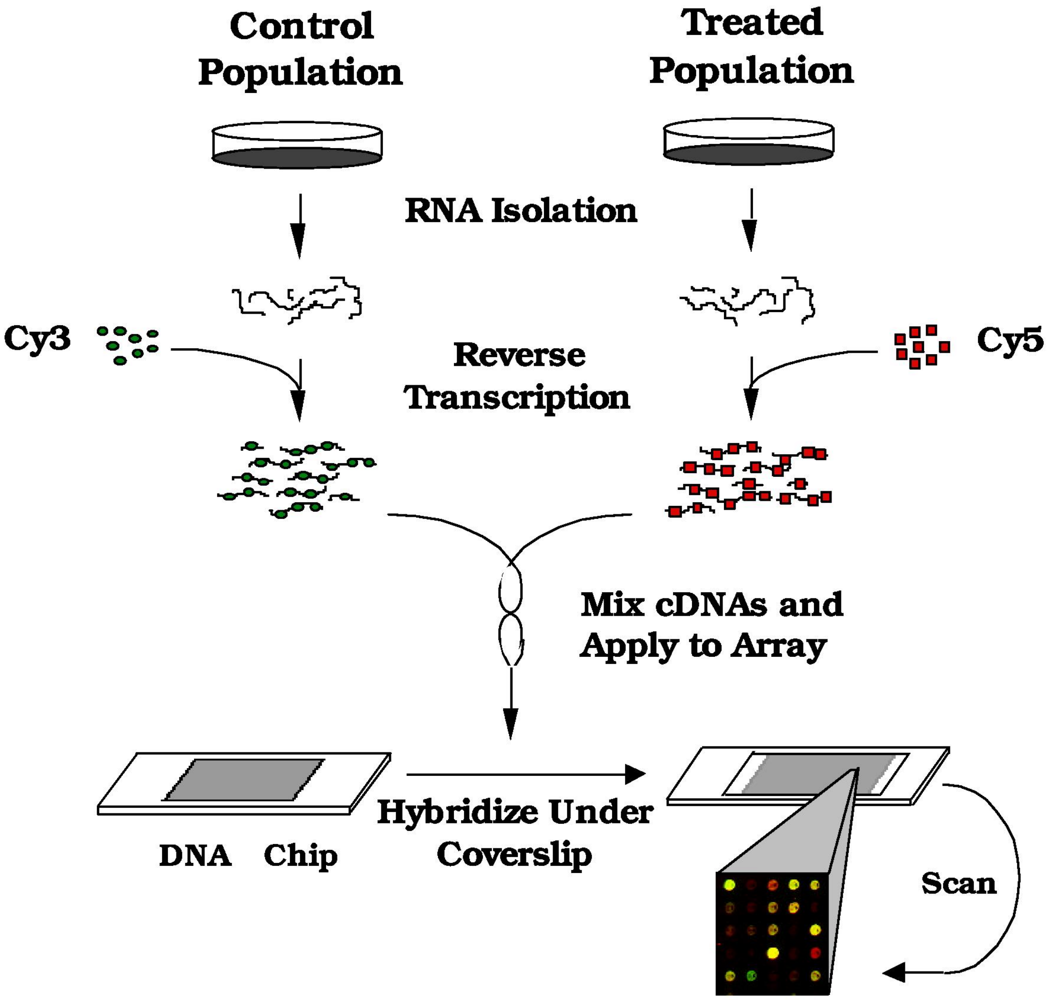
Numerical operations (like this square-root transform) can make certain features more visible. Transformations must only be used for visualization, since they would distort the quantifications.

Nylon cDNA Microarrays

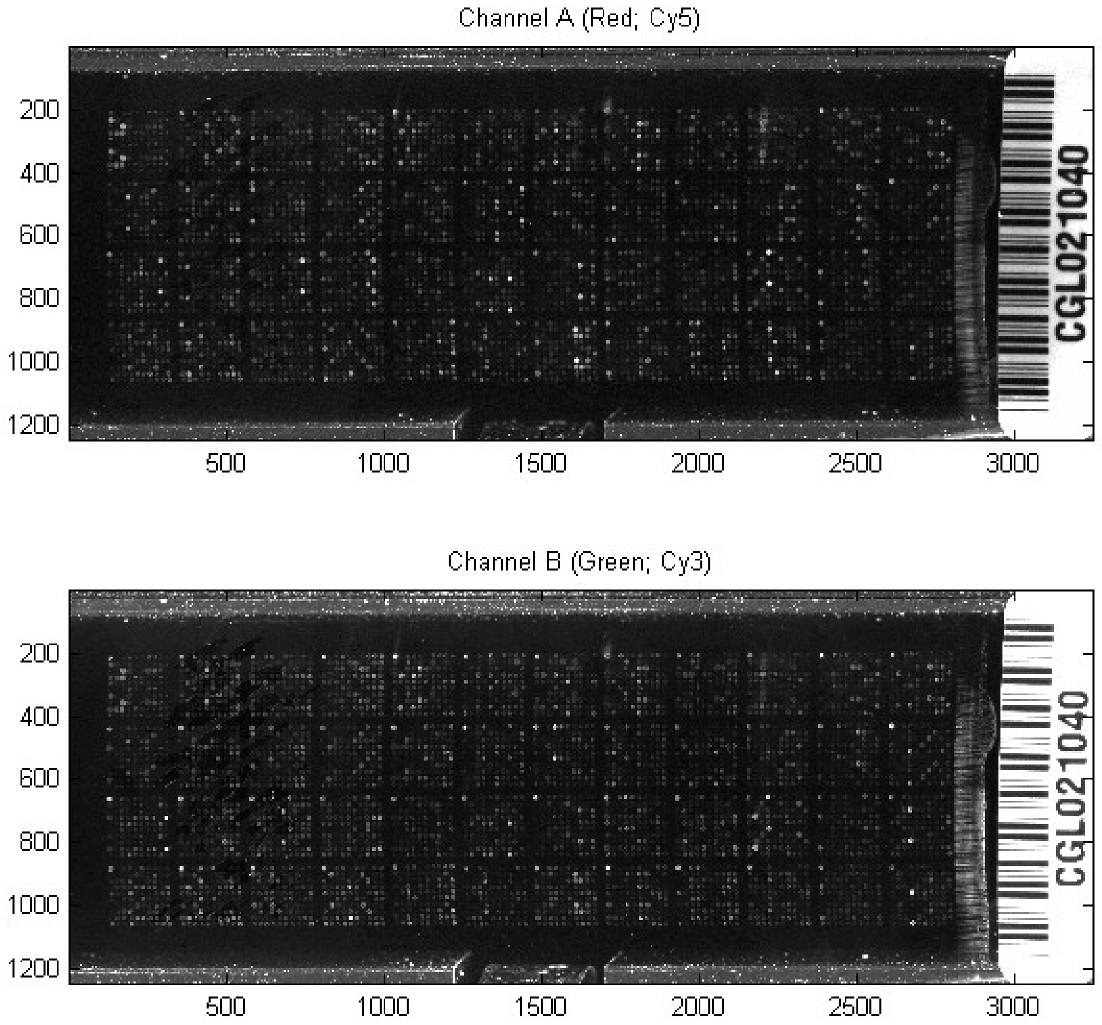


This is a close-up of the same microarray. Note the general blurriness, along with the blotchy artifact along the left edge.

Overview: Two-color Spotted Microarrays



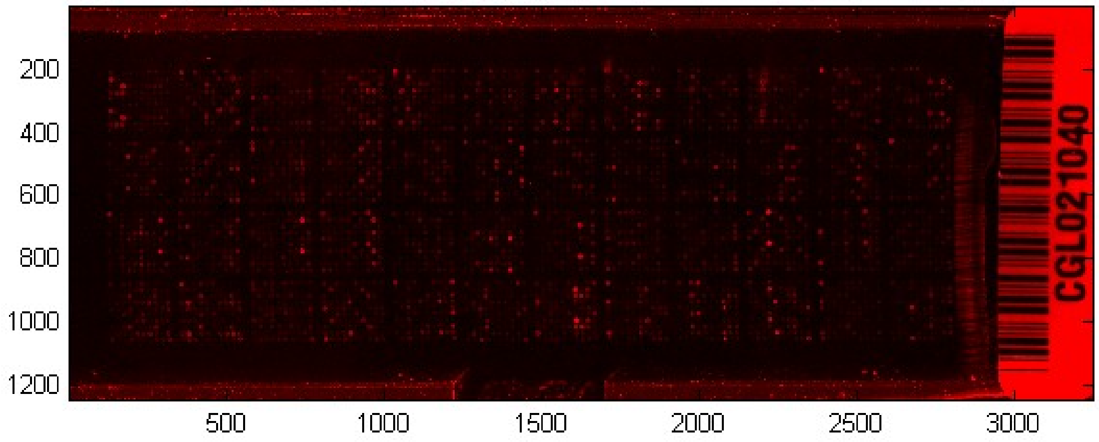
Two-color Spotted Microarrays


Channel A (Red; Cy5)

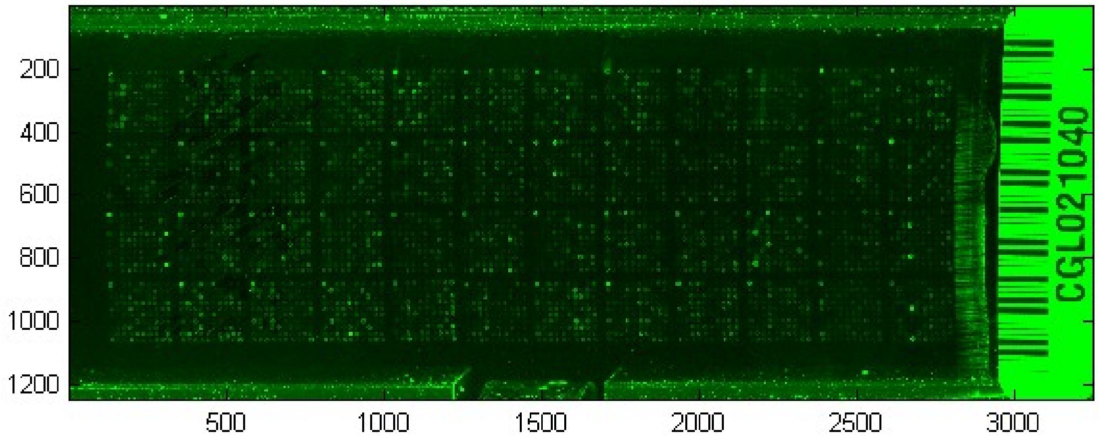
Channel B (Green; Cy3)

Two-color Spotted Microarrays

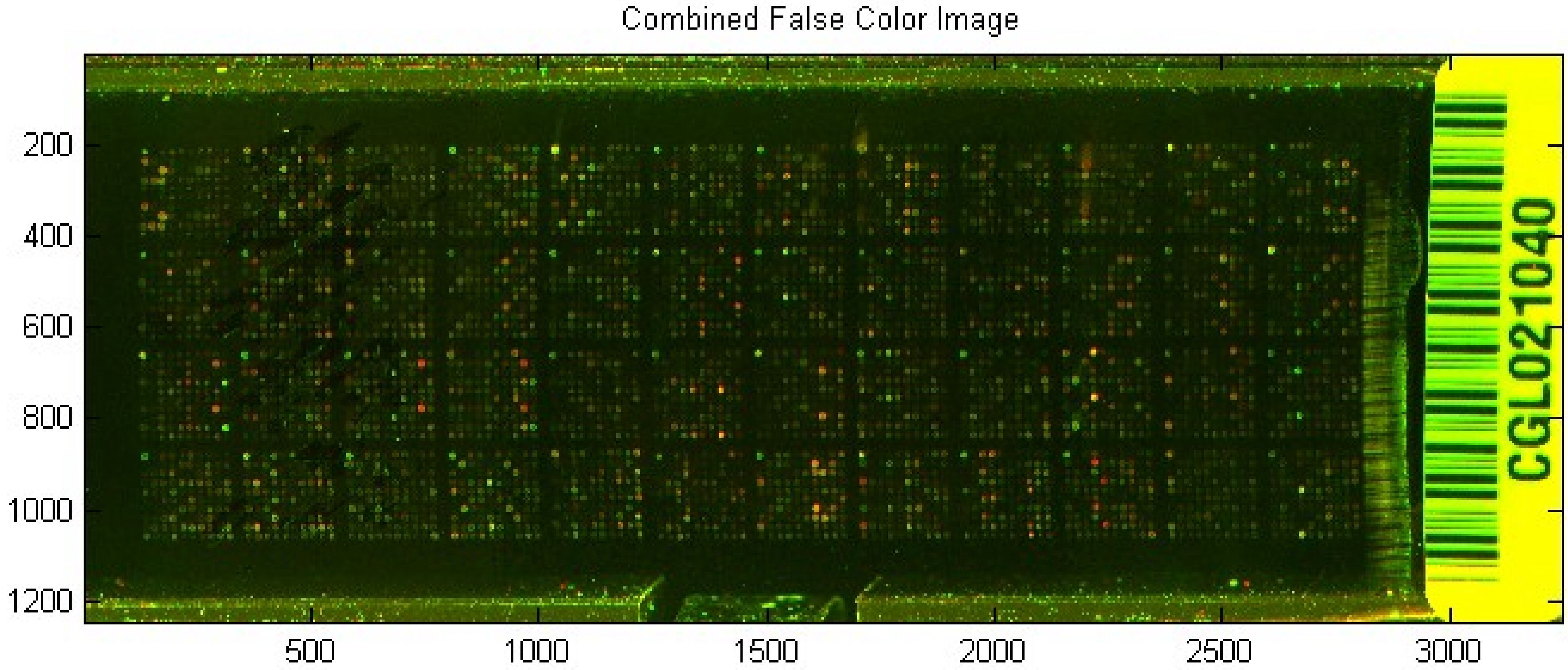
Channel A (Red; Cy5)



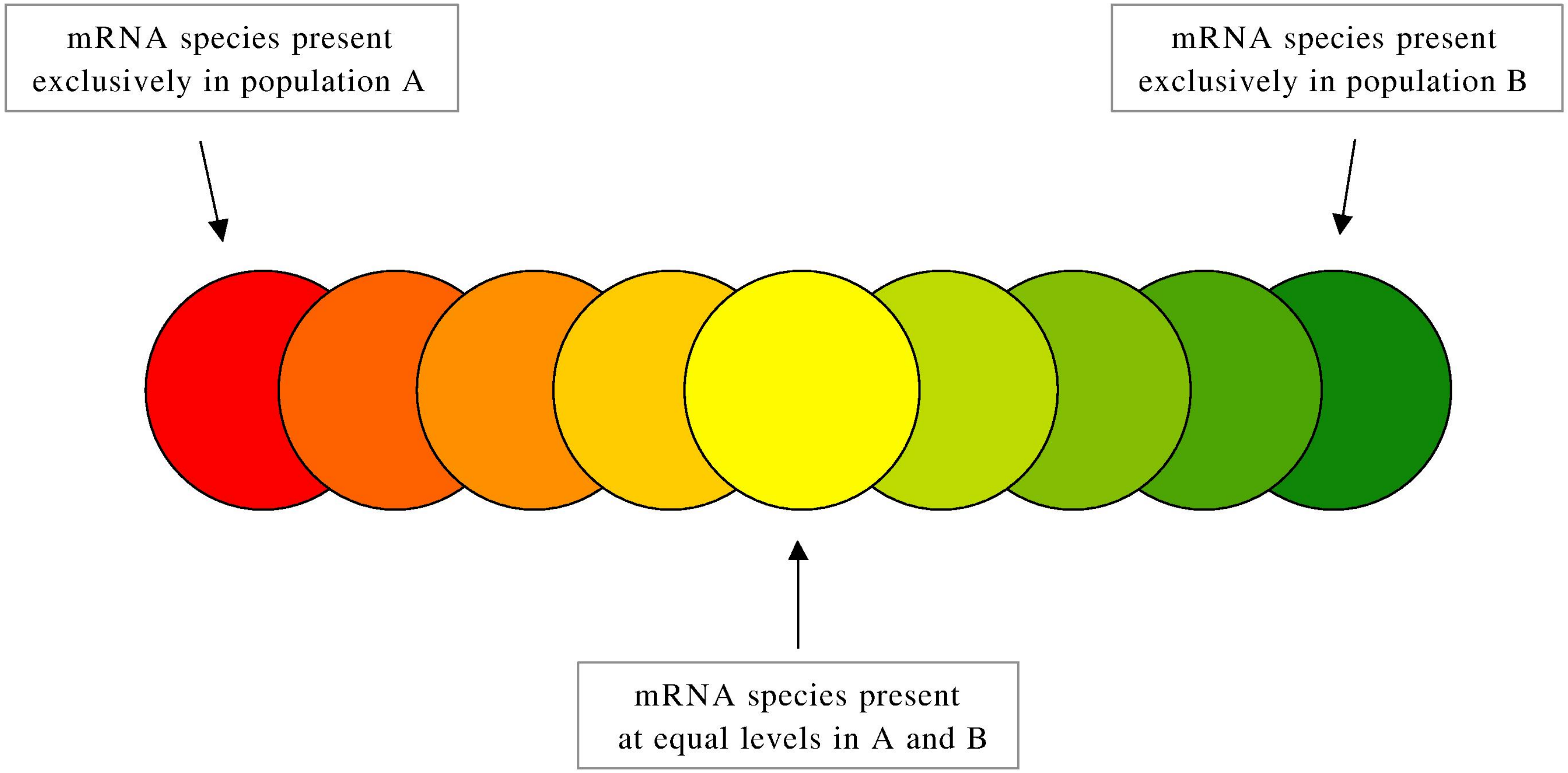
Channel B (Green; Cy3)



Two-color Spotted Microarrays



Two-color Spotted Microarrays



mRNA species present exclusively in population A

mRNA species present exclusively in population B

mRNA species present at equal levels in A and B

The Images Are The Data

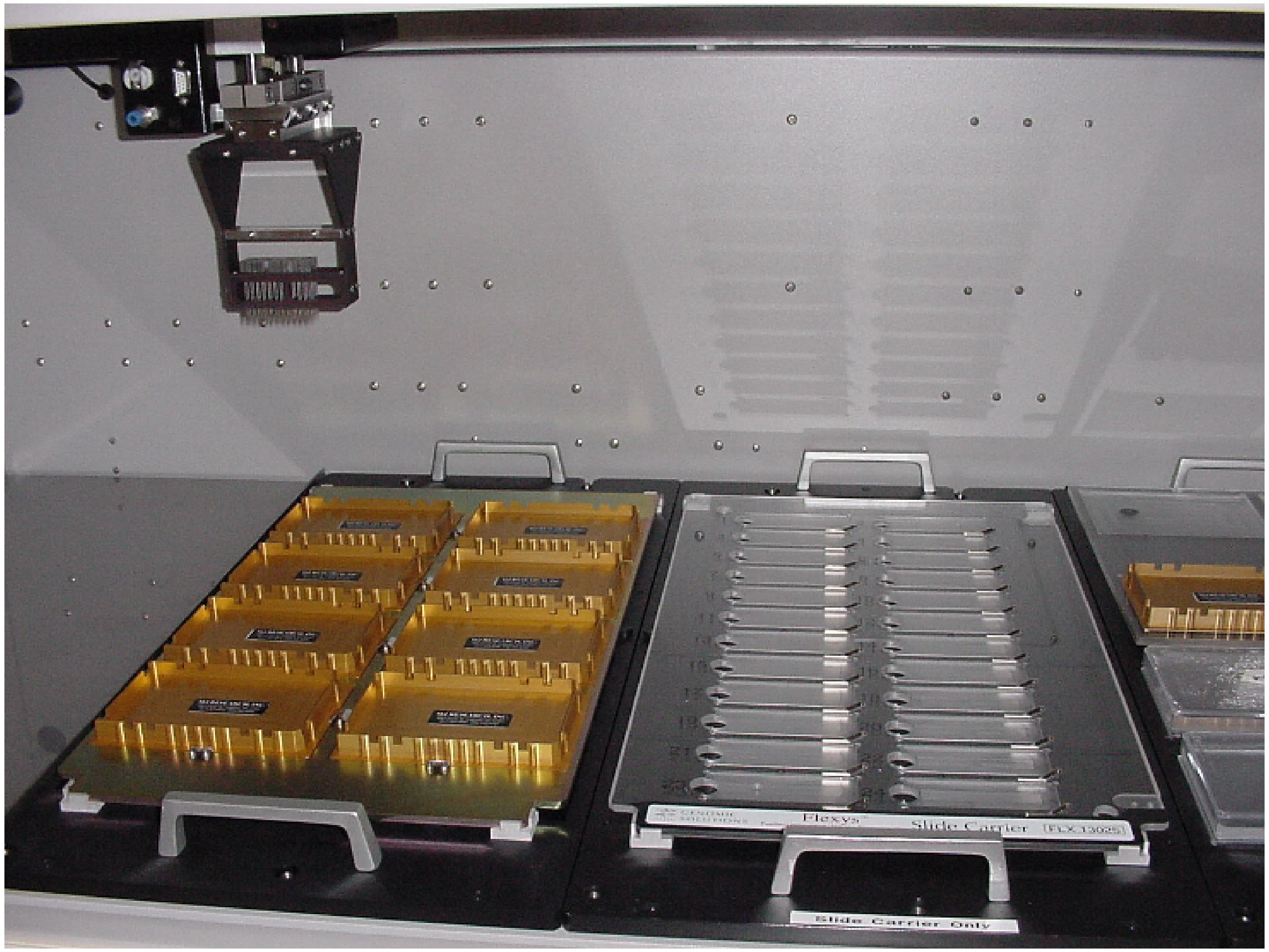
A common feature of all microarray platforms is that the primary data produced by an experiment is in the form of a gray-scale image.

With Affymetrix arrays, the raw image is stored in a DAT file. The precision of the manufacturing process and the alternating border of dark and bright spots makes it fairly straightforward to find and quantify features.

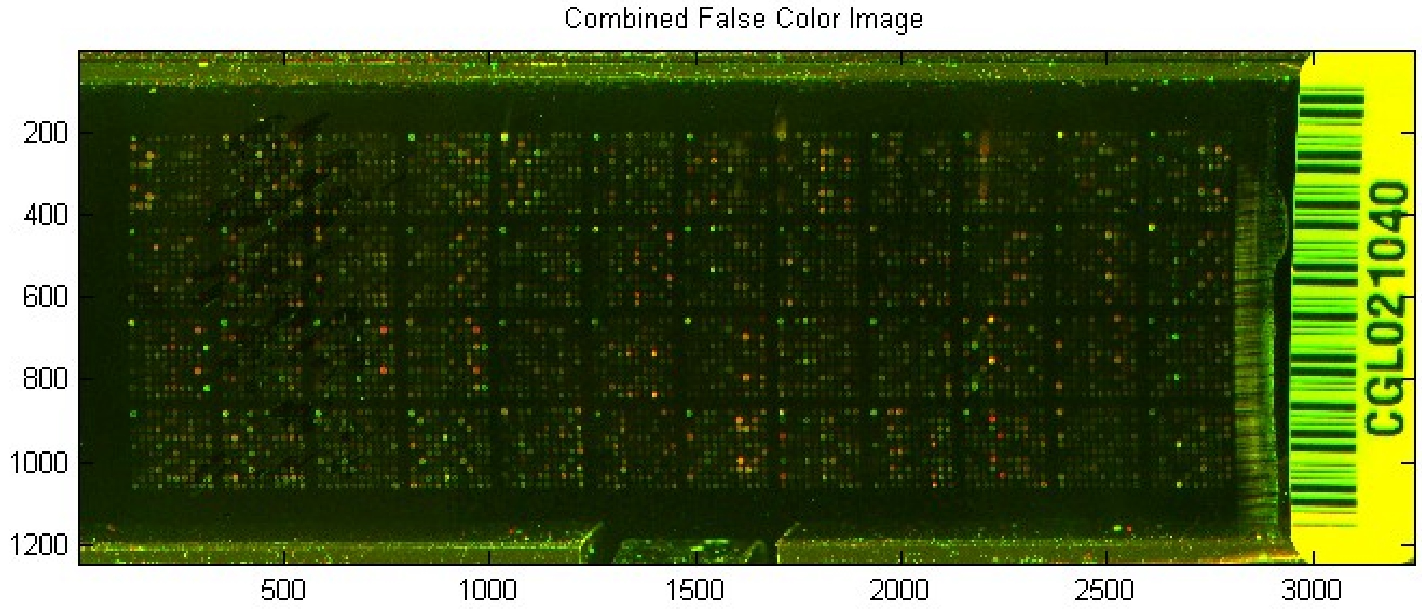
Mechanical variation in the robotic production of nylon or glass arrays, however, makes finding and quantifying features a more complicated procedure. Extra complications also arise from the pairing of two samples (red and green) in the same hybridization.

Arrayer robot

Glass microarrays are produced by robotically spotting cDNA or long oligos (60- or 70-mers) on glass microscope slides.



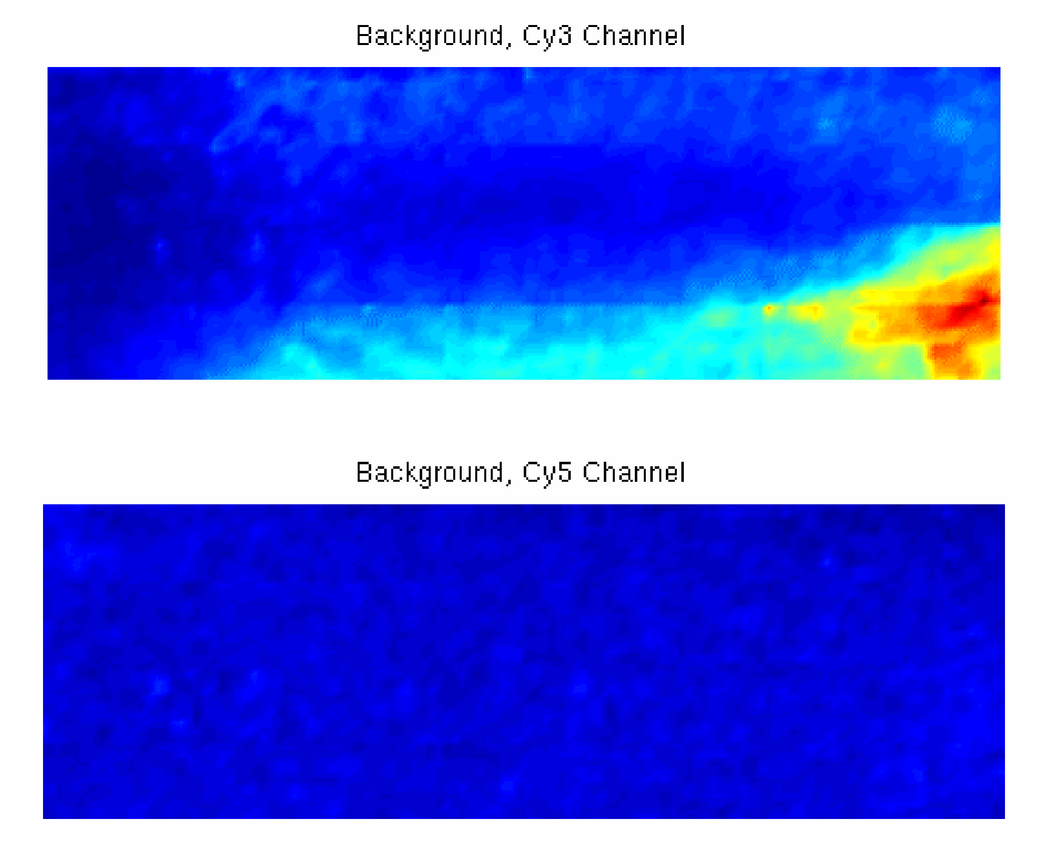
Robotic pins impose a grid substructure



This slide was produced by a robot using 48 pins, laid out in a 4×12 rectangle. Each 10×10 subgrid was produced by a separate pin.

Note: Flaws in a pin (e.g., blunt tip) can cause one subgrid to be systematically different from the others.

Trends in the background

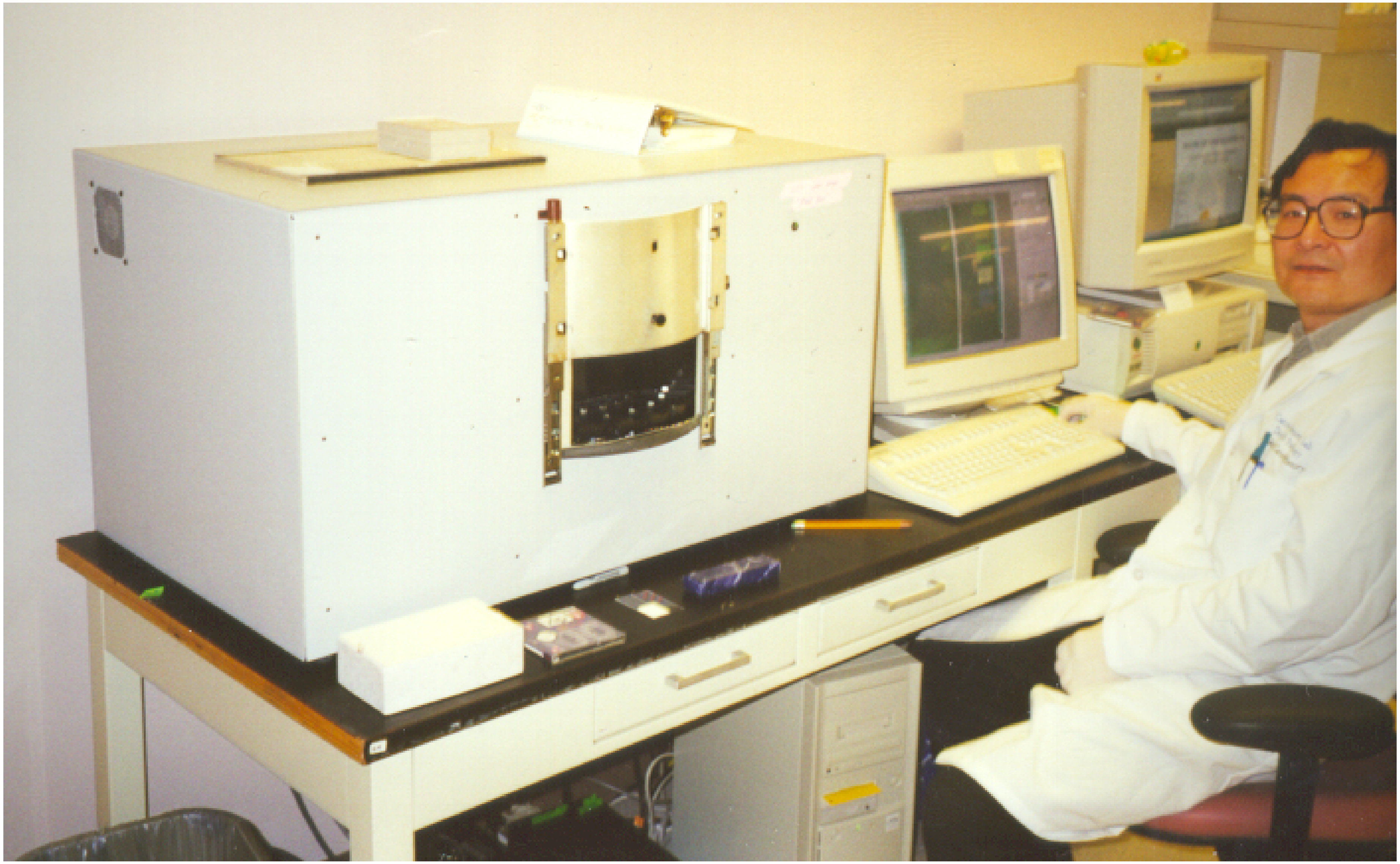


Background, Cy3 Channel

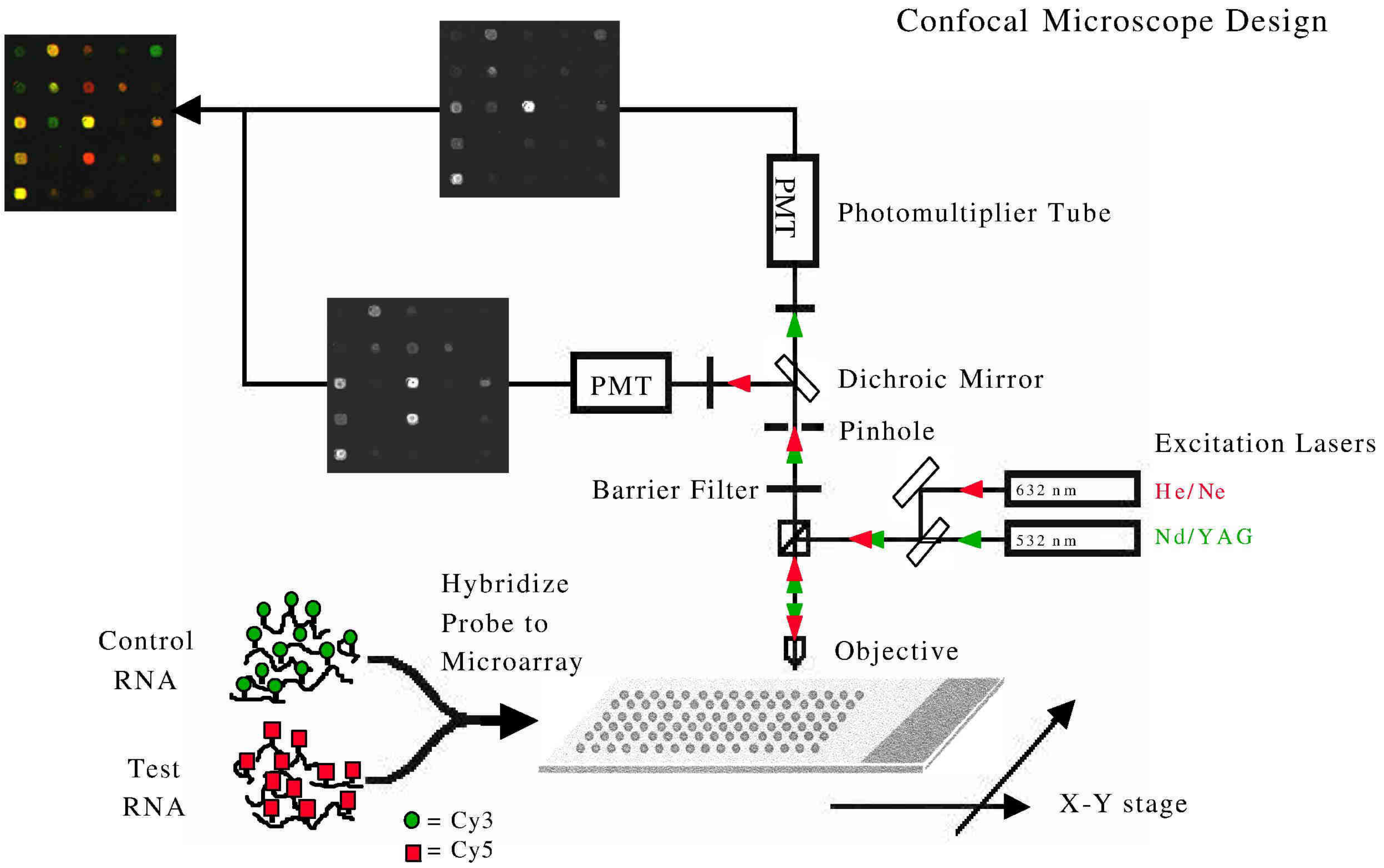
Background, Cy5 Channel

Systematic differences in grids can create spatially coherent artifacts. These can also be caused by incomplete mixing of the hybridization solution, often revealing itself in the background.

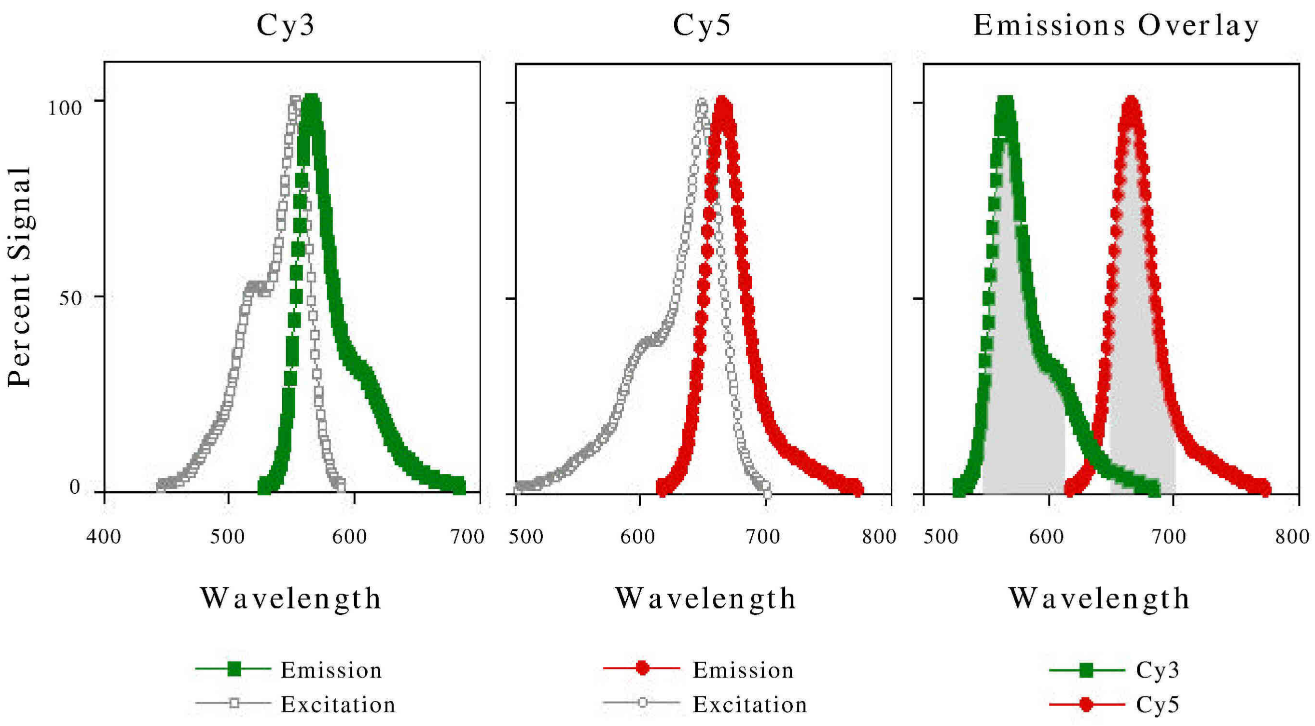
Scanners are boring beige boxes...



Scanning schematic



Fluorescent Emission Properties



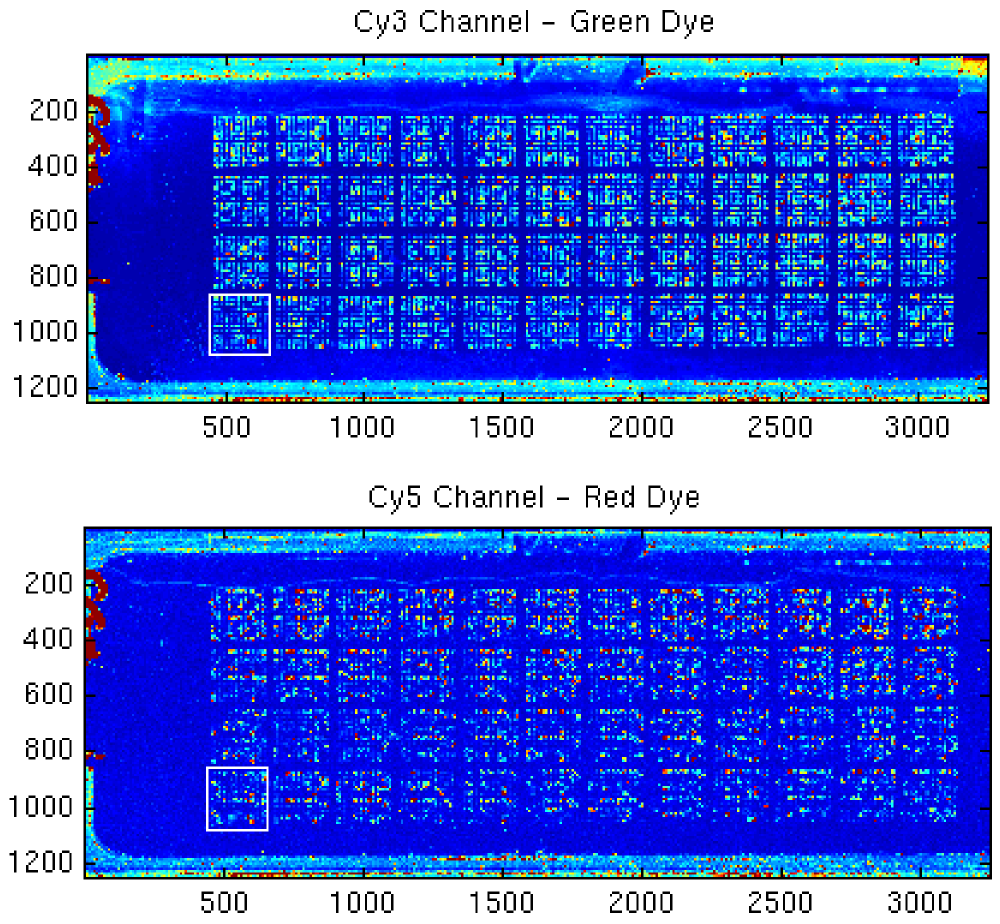
Fluorescent dyes (Cy3 = green, wavelength 532; Cy5 = red, wavelength 635) absorb light at a specific excitation wavelength, and emit light at a specific larger wavelength.

Dye effects

The basic difficulty in making sense of glass microarray data is deciding how to combine the information from the two channels. The two dyes have different chemical properties; they may be incorporated into genes at different rates. They may also fluoresce at different rates.

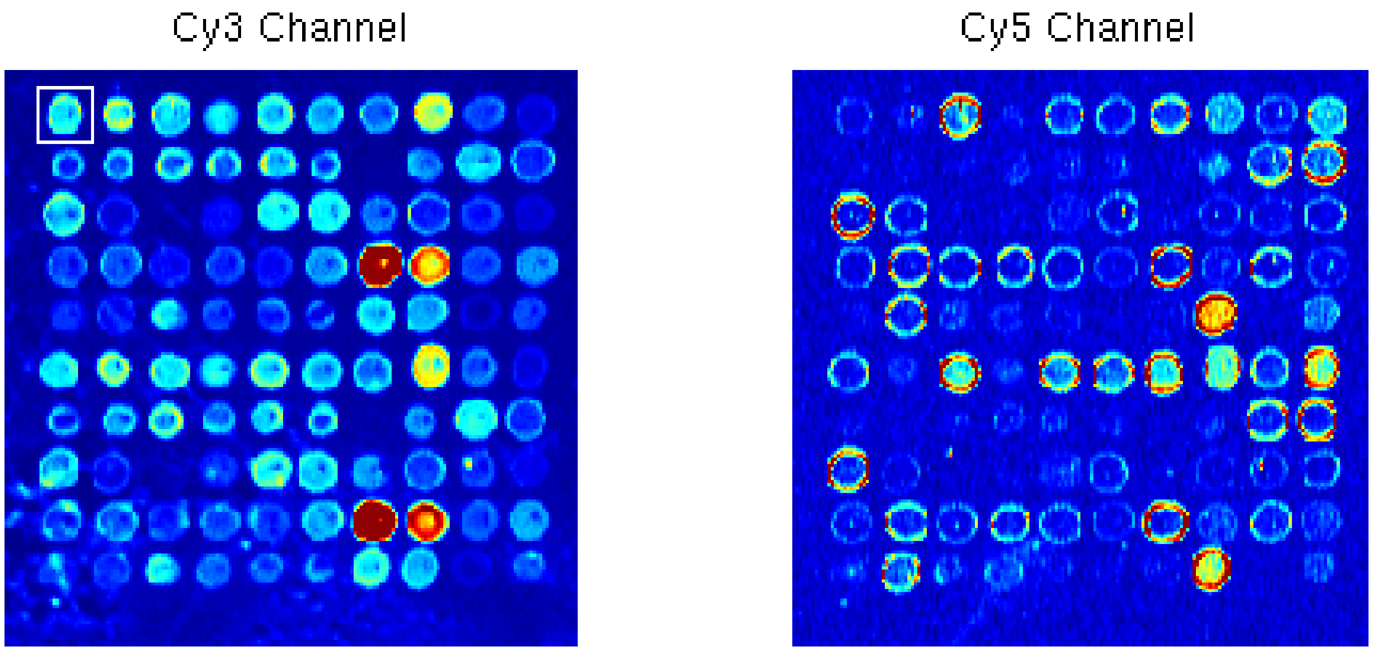
These differences can lead to array-wide differences in intensity and to gene-specific differences related to the DNA sequence of the target genes. These differences have implications for the downstream analysis, and will therefore affect how we think about designing microarray experiments.

A closer look at a microarray image



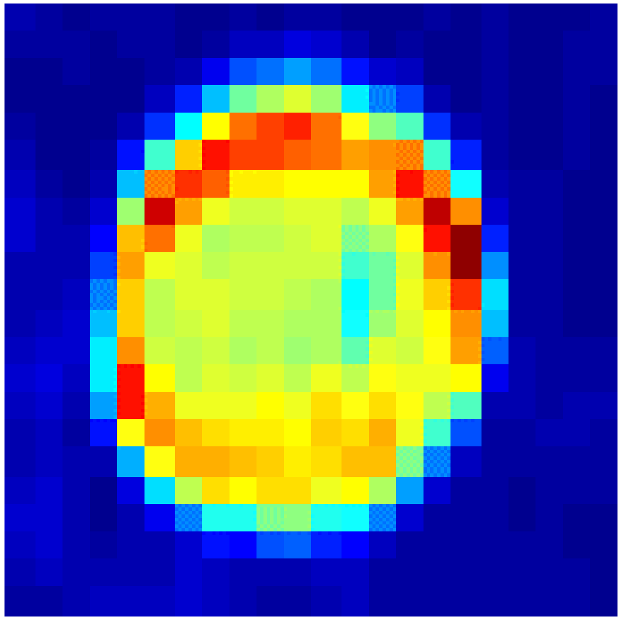
This is a fairly old microarray from M.D. Anderson; it doesn’t yet have a barcode attached.

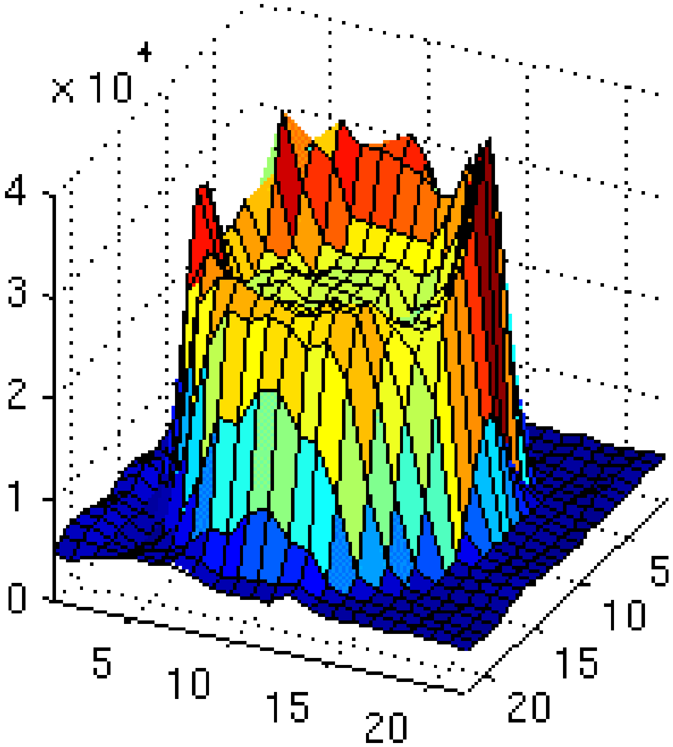
A closer look at one subgrid



These arrays were printed with duplicate spots. The top 5 rows are duplicated in the bottom 5 rows. Note the "ring" or "donut" effect that is evident at many spots, especially in Cy5. [Deegan et al (1997). Capillary flow as the cause of ring stains from dried liquid drops. Nature, v.389, p.827-9.]

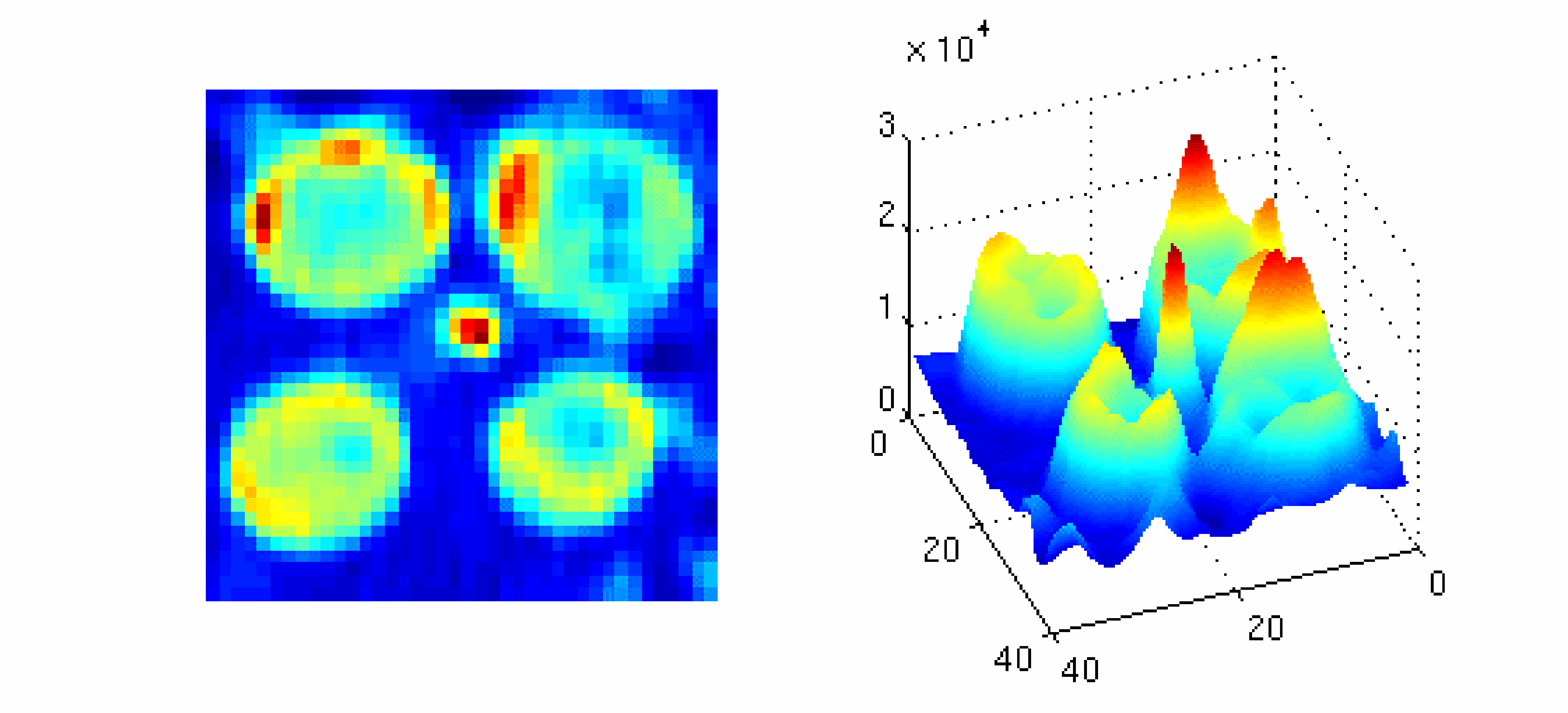
A closer look at one spot





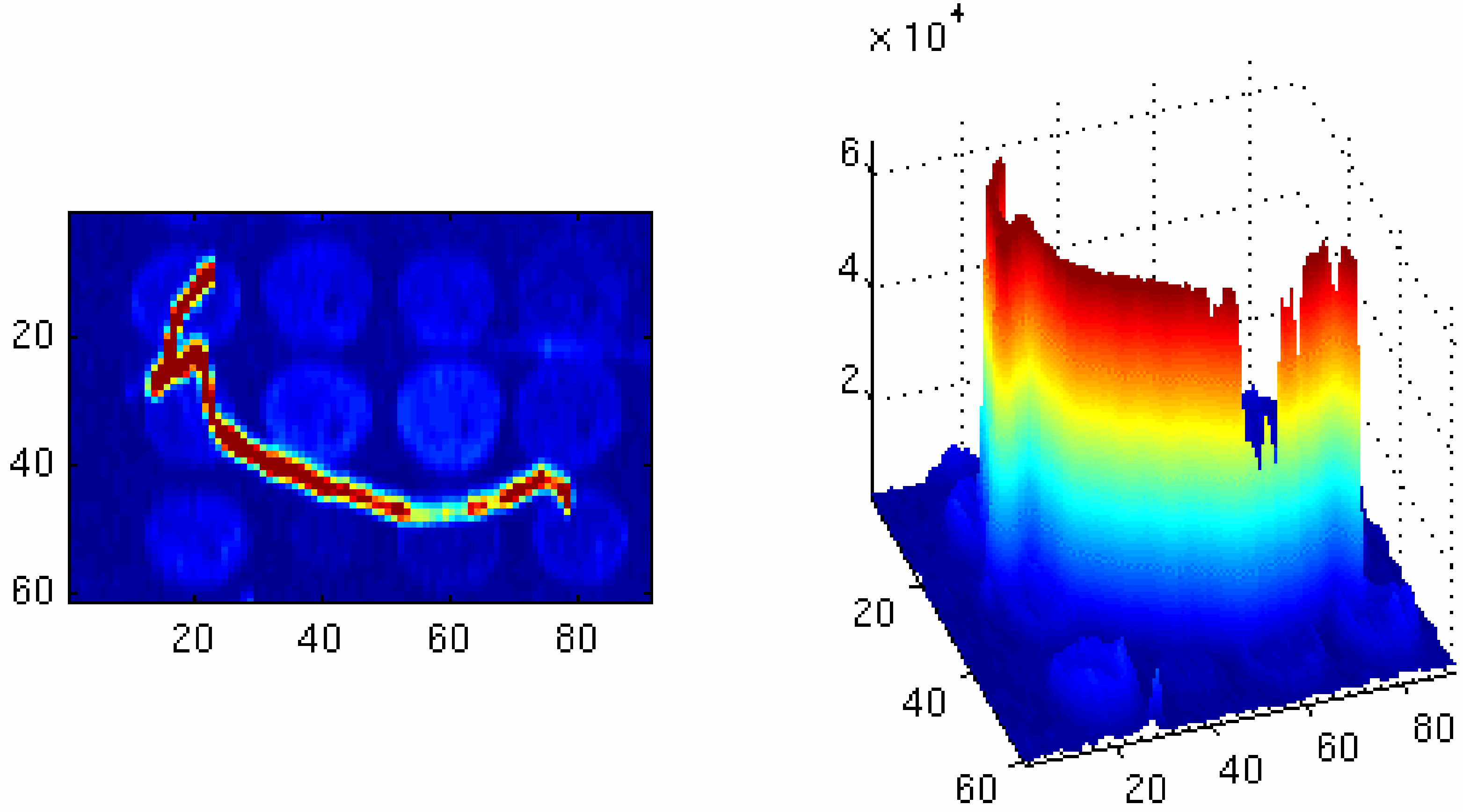
Notes: (1) The spot is not perfectly uniform; higher edges are evident even in bright spots. (2) The background is not uniform. (3) The spot appears to be contained in a box with 22×22 pixels.

A dust speck



Automatic processing algorithms to locate the spots have to contend with large dust specks; quantification algorithms can be affected by smaller specks that overlap true spots.

A fiber



Why do you think the pixel values are uniform along the fiber?

Summary: glass microarrays

We've seen a number of potential difficulties with analyzing the images from glass microarrays.

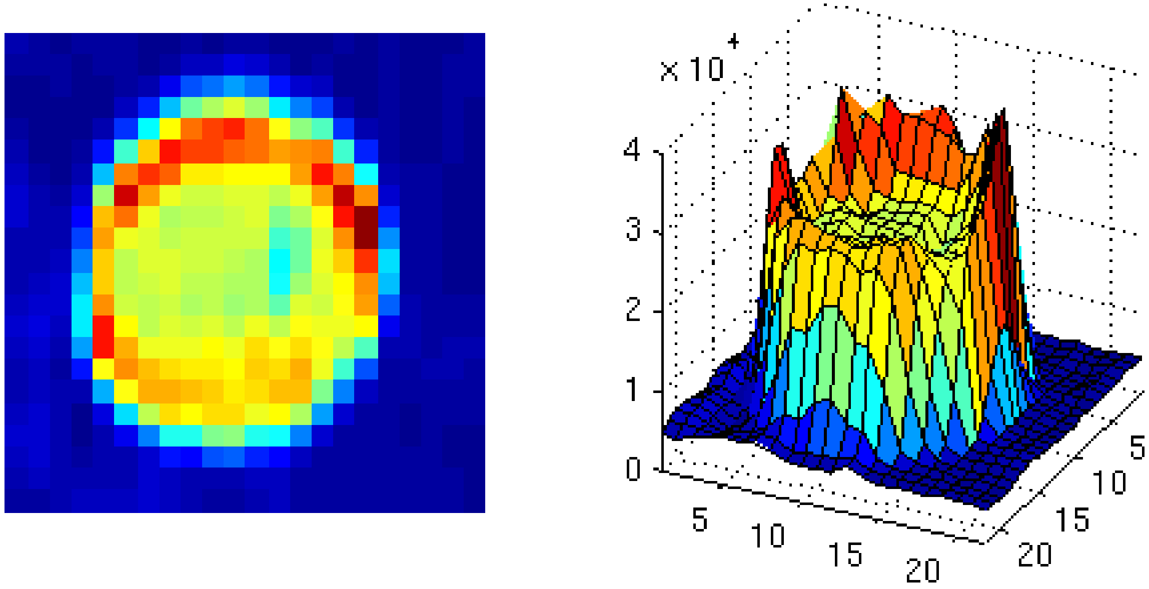
- Artifacts like dust specks, fibers, or water spots can cause small-scale problems with spot-finding or quantification.

- Differences in pins can cause systematic biases on the scale of subgrids.

- Channel differences, either directly related to chemical propoerties of the dyes or differences in laser intensity, can introduce systematic distortions.

- Insufficient mixing of the hybridization solution can cause large-scale differences in background and signal intensity.

The Next Goal: Quantifying a Spot



TIFF Files: The Good

Almost all cDNA microarray images are 16-bit TIFF files.

This is a fairly standard file format, and most image viewing software will read TIFF files. . . However

The TIFF standard is highly flexible and can even be customized. In addition to the basic parameters required for all images, you can add your own “tags” to the files to guide processing.

TIFF Files: The Bad

Flexibility can be unclear. The defined behavior on encountering an unknown tag is to skip that field entirely.

The above can bite you. (More shortly.)

Flexibility in terms of placement of tags in the file and possible inclusion of internal compression makes it harder to find (or write) freeware scripts; some of the compression algorithms are not public domain.

R does not have a built-in function for reading TIFF files.

Dilution: A Cautionary Tale

STORM phosphorimagers can be used to produce image files from radiolabeled nylon membranes.

The STORM counts are advertised to go to 100,000.

A 16-bit register holds counts up to 65,535 ($2^{16} - 1$).

How do they do it? They record the square root of every count! If you don’t know this, your assessments of the amount of change will be quite off.

A Quantification File

A typical entry:

```
Spot labels,     AR VOL - Levels x mm2,
A - 1 : a - 1,   1585.1254,

% Replaced (AR VOL), SD - Levels
0.000,               13869.45

Pos X - mm,  Pos Y - mm,  Area - mm2
11.278,      21.346,      0.083

Bkgd,     sARVOL,    S/N,     Flag
396.710,  1188.416,  33.006,  0
```

What do these numbers mean?

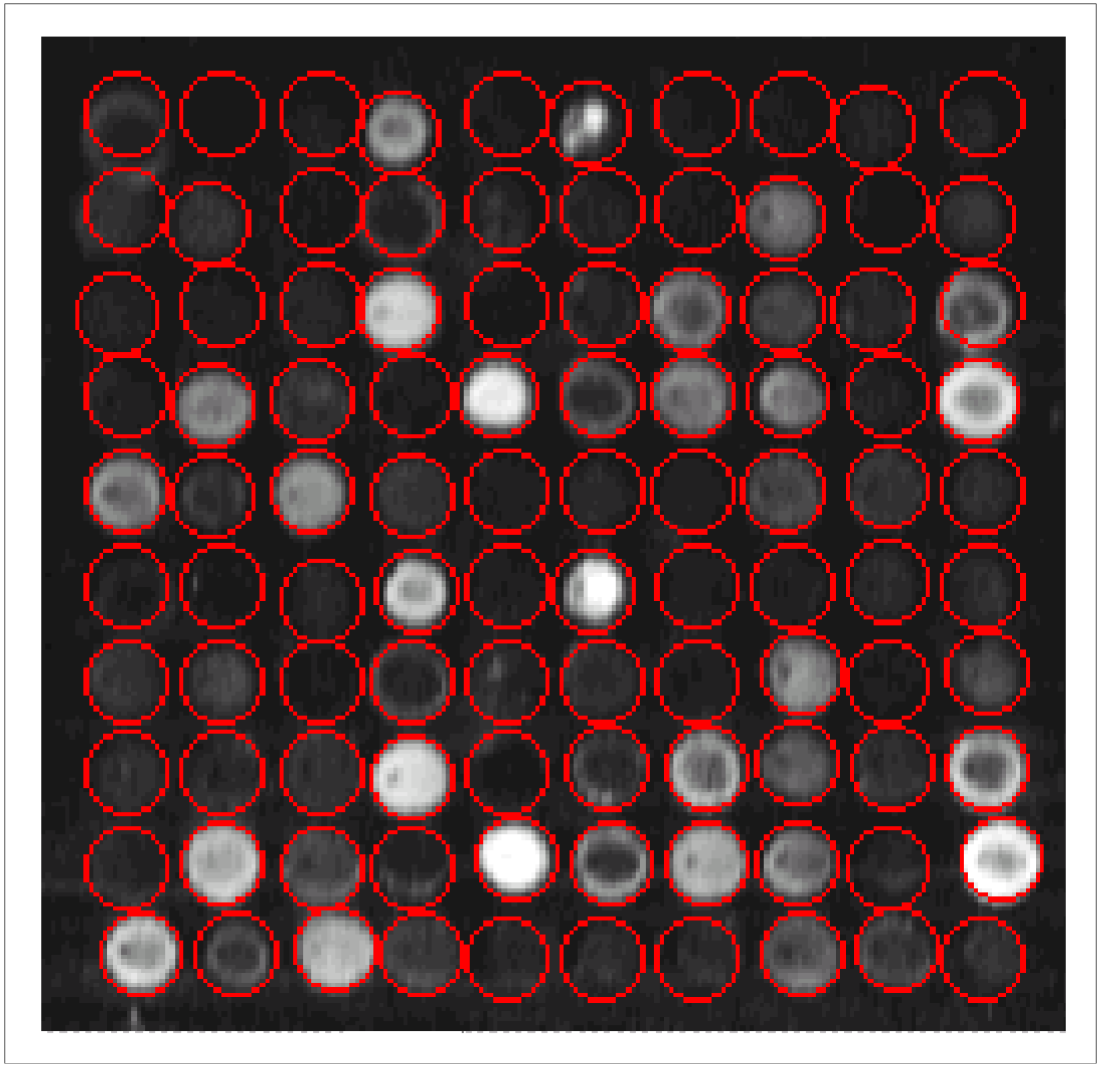
Quantification has three parts

Registration: locating the centers of the spots

Segmentation: deciding which pixels around the center actually belong to the spot

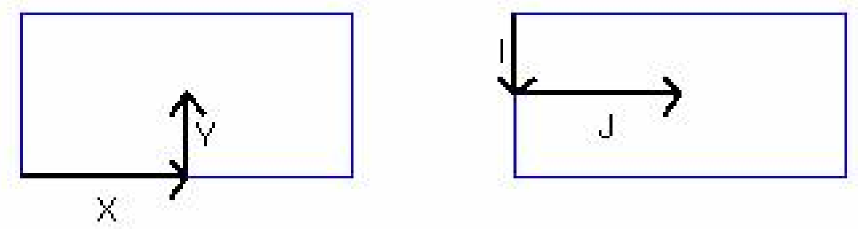
Quantification: summarizing the numerical pixel values in the spot (foreground) and around it (background).

Quantification in action



Position

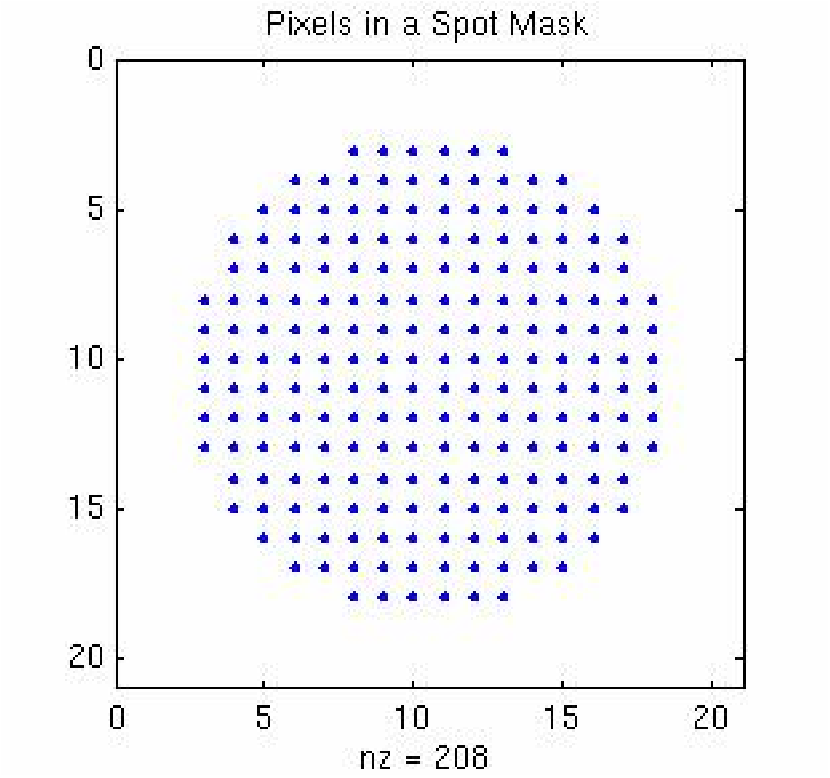
Pos X - mm : 11.278



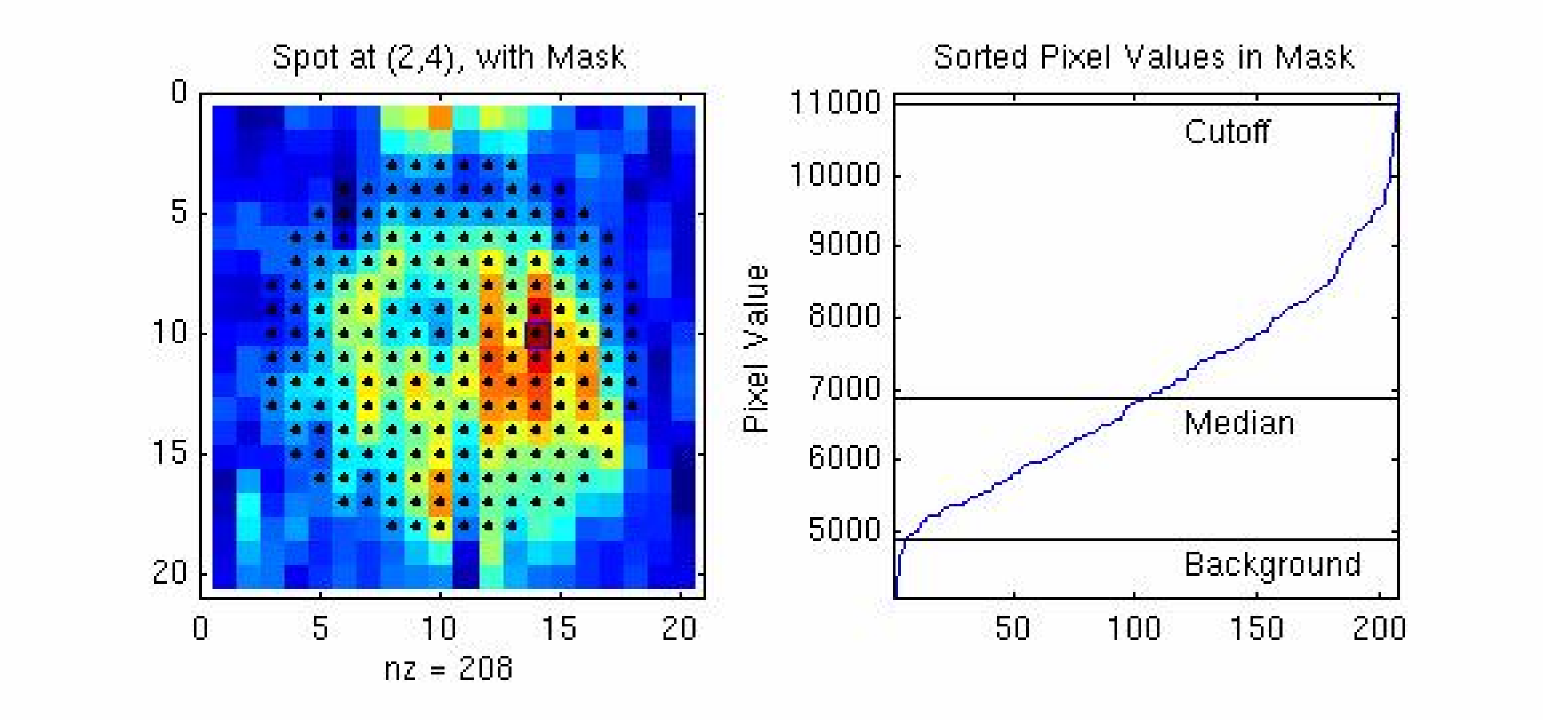
specifying the center of a spot in terms of the slide (mm, from lower left) and image (row and column, from upper left).

Volume: Summing the pixels in the circle

Having identified the center of the spot, we add up all the pixels in a circular region here (the “mask”). The size of the mask is dictated by the physical spot size.



What do the pixel intensities look like?



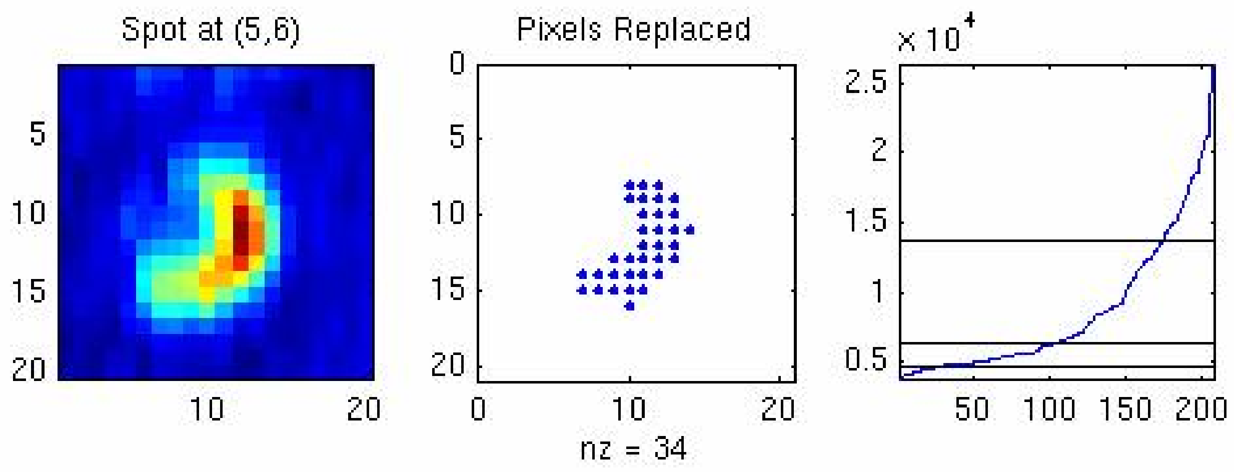
Not quite a perfect fit, but ok.

What is AR Vol?

AR Vol = “Artifact Removed” Volume. As most artifacts are of high intensity, we omit those whose intensity is very far above the central (median) value that is seen.

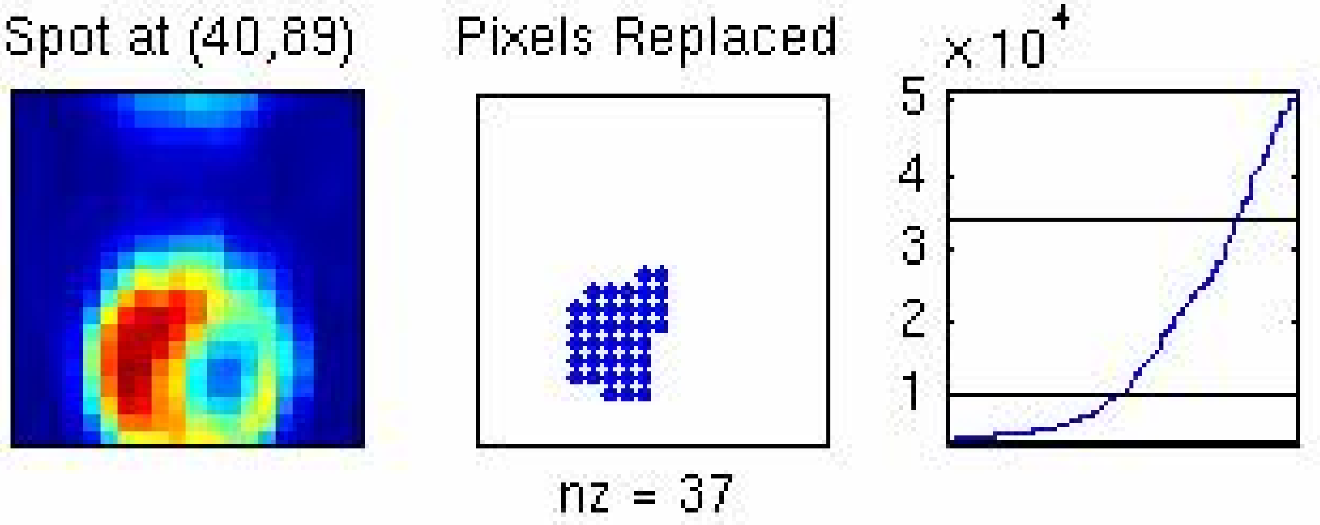
This presumes that the spots are (a) even, and (b) of roughly equal size.

Does AR Vol Work?



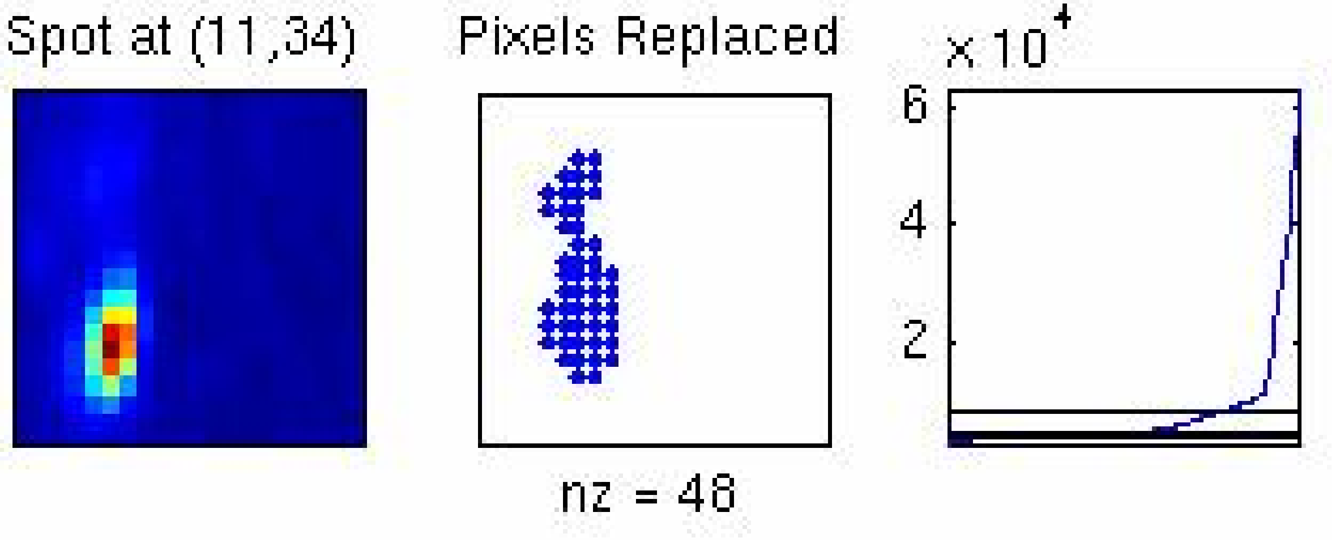
Not always...

Does AR Vol Work?



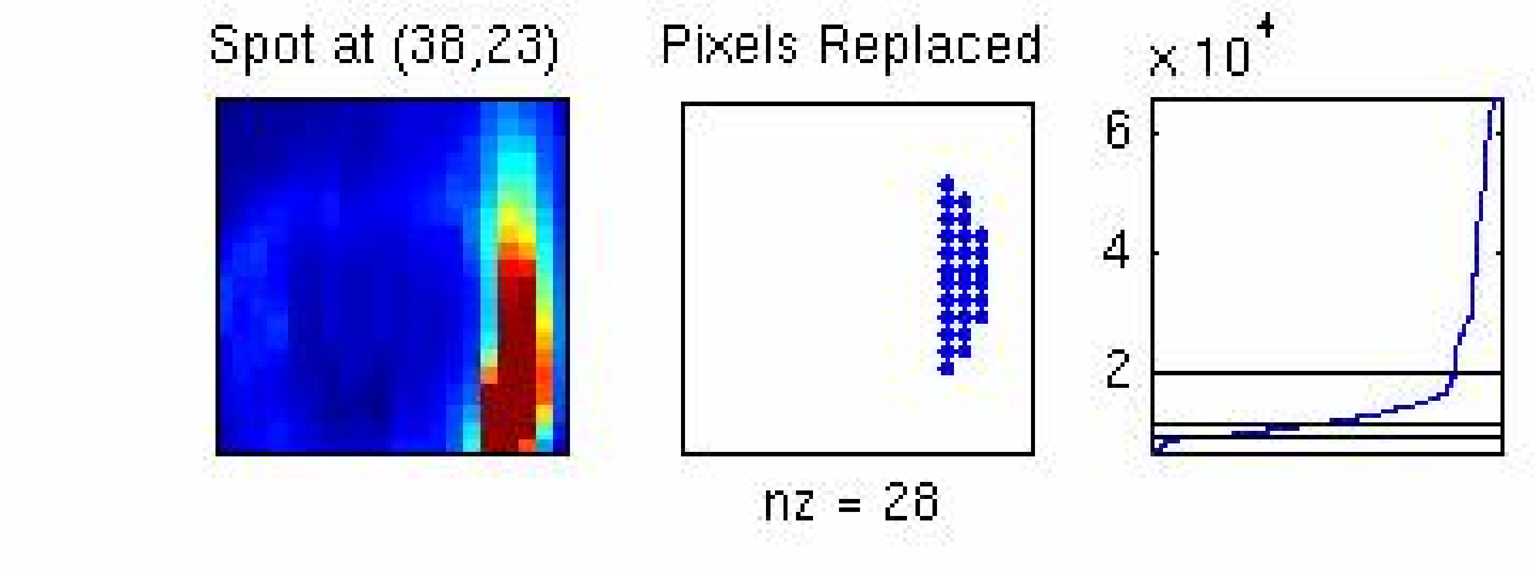
Not always...

Does AR Vol Work?



But sometimes!

Does AR Vol Work?



But sometimes!

Are there other ways of dealing with bugs?

The simplest and best is to use replicates, either within the array itself, or across multiple arrays.

Replicates let us check that the differences are “large” relative to the scale of “Noise”.

What other metrics could we use?

mean?

median (or some other percentile)?

why a circle?

There are downsides to a fixed shape. We may miss some parts of the spot, and include others that “aren’t there”. (segmentation)

What about Background?

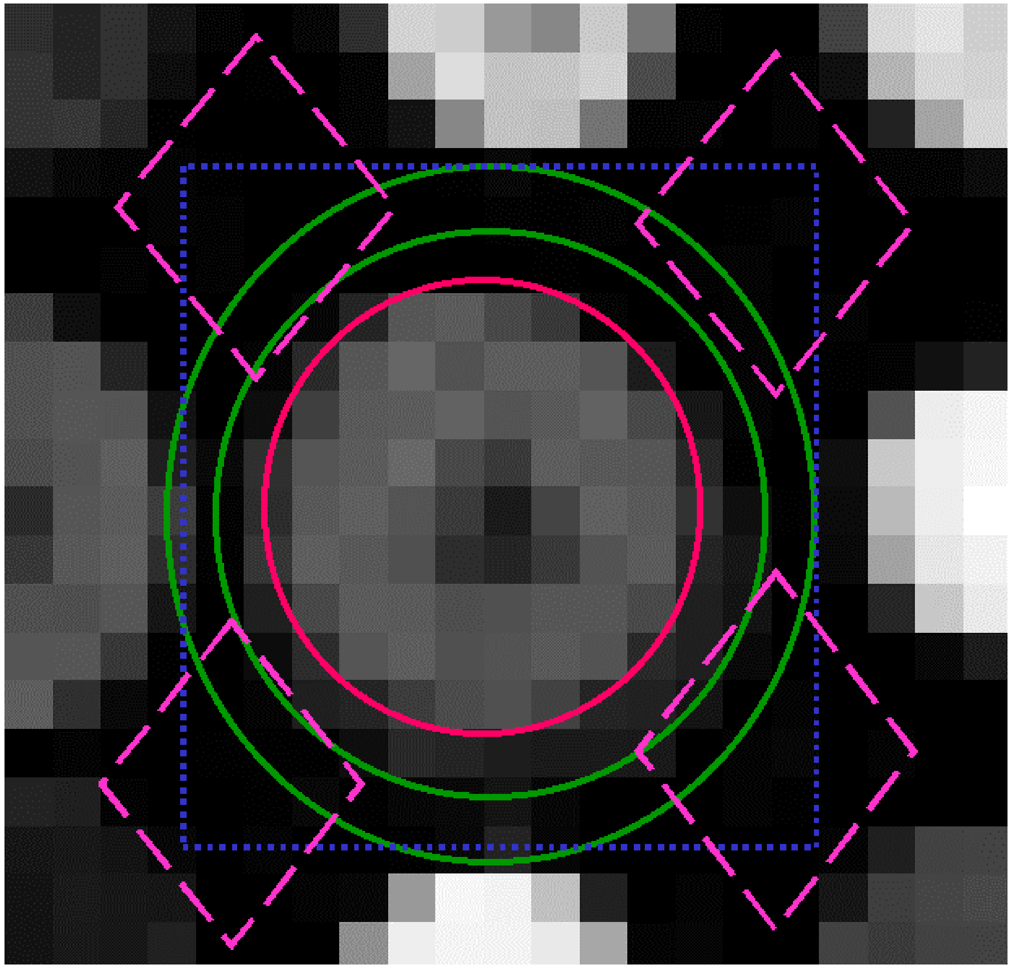
What is it? The pixel intensity in those parts of the image where nothing has been spotted.

How do we count it? Typically by averaging the intensities in some “non-spot” pixels close to the spot center.

What do we do with it? Subtract it off from all pixels inside the spot(?).

Subtracting background is very important when a fixed mask is used.

What about Background?



Which pixels do we use? (Yang et al, JCGS, 2001)

Putting Channels Together: Log Ratios

Why is this a natural scale?

Lots of biological stuff works in terms of fold changes.

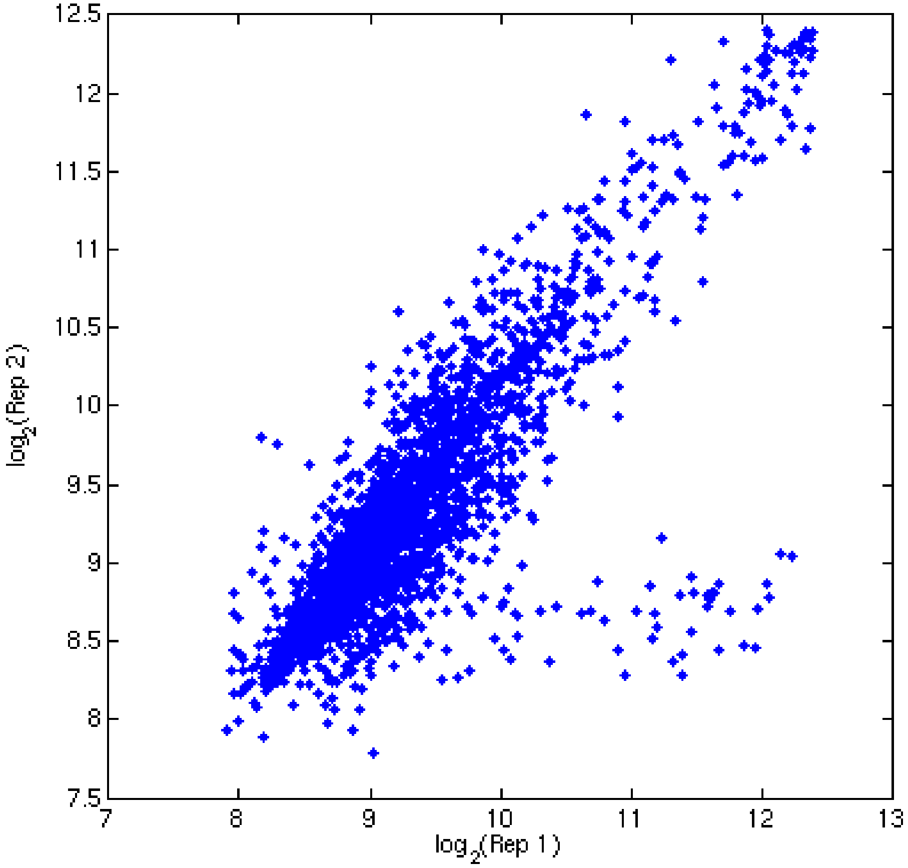
Why log?

Fold changes of 2 and 1/2 are of equal magnitude, but different sign

Why ratios?

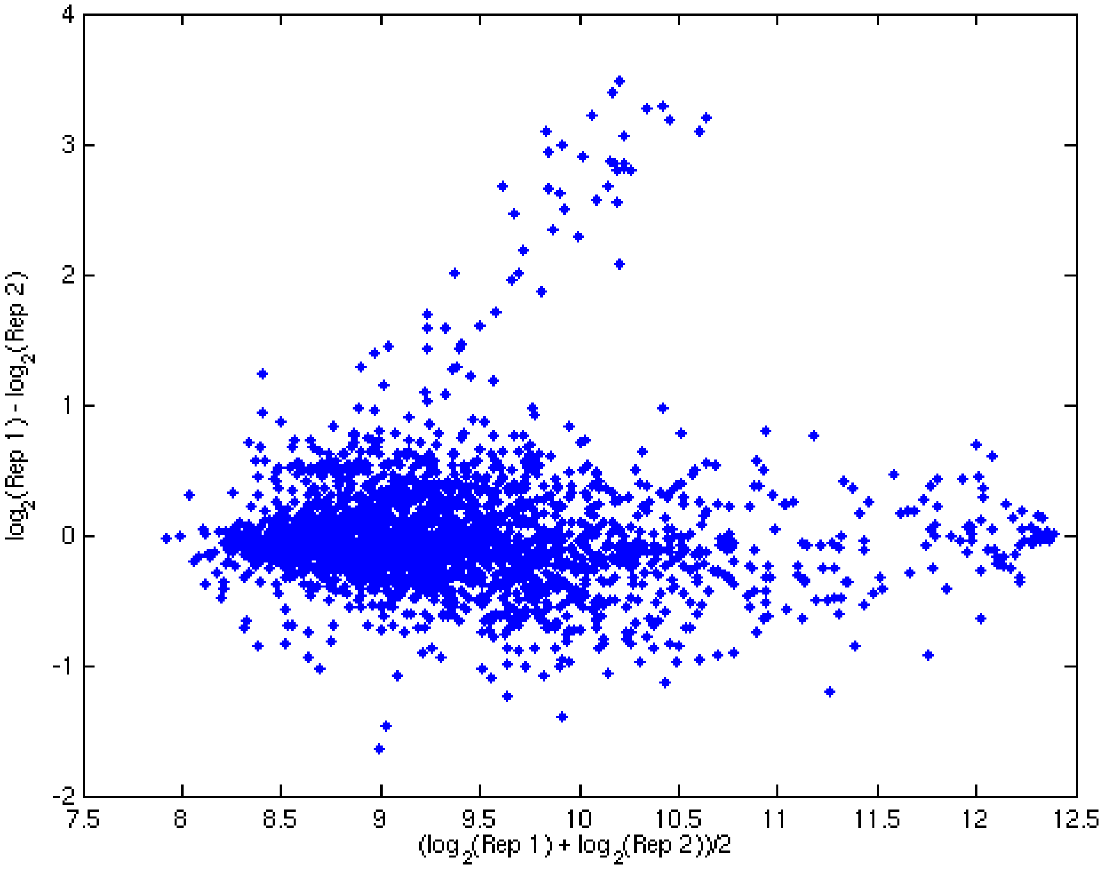
We'll look at that below.

Checking the Data: Replicate Spots



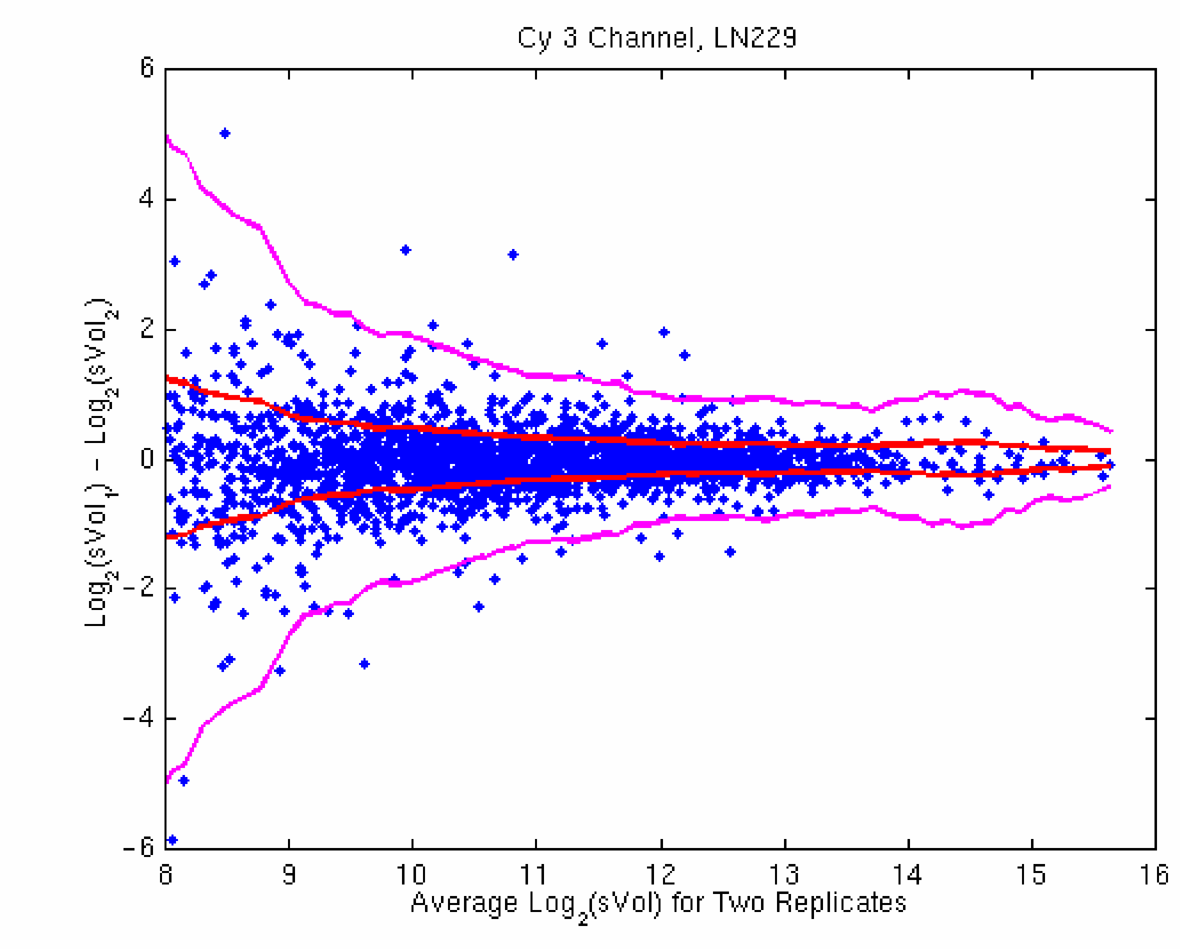
Log scale; do things line up?

A Better View: M-A Plots



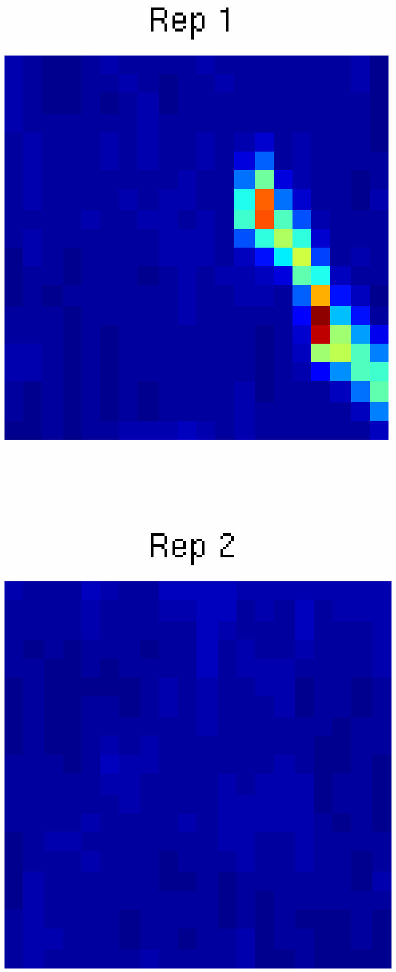
rotate things by 45 degrees so that they're easier to see.

Subtracting Background, with Replicates

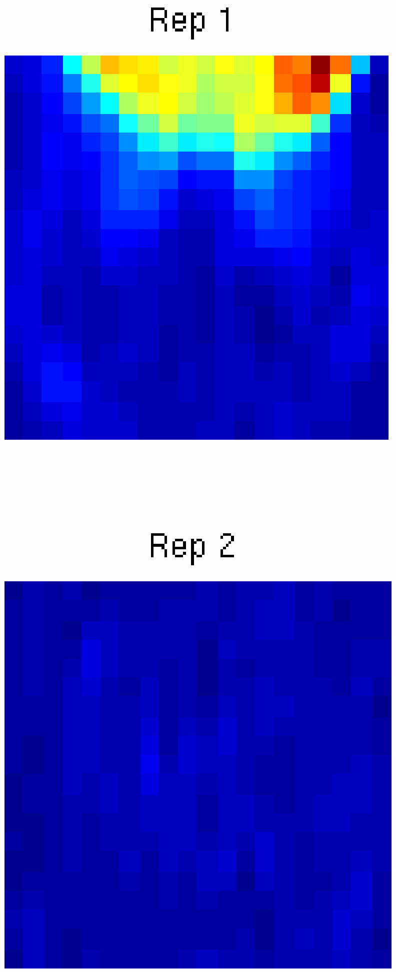


Negative values, thresholding, and variability

Flagged Spot 1

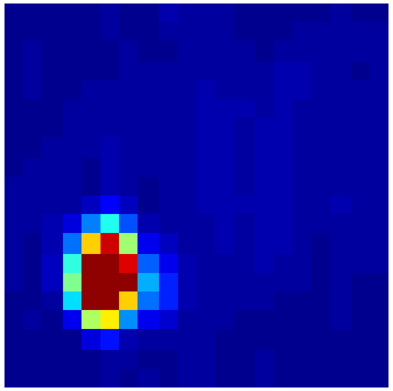


Flagged Spot 2

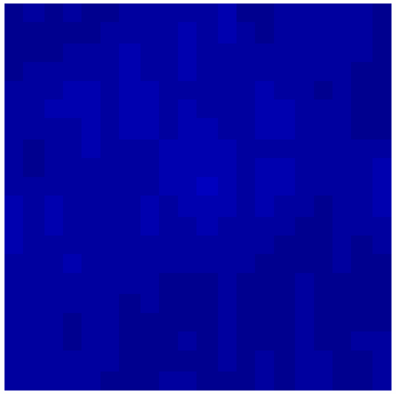


Flagged Spot 3

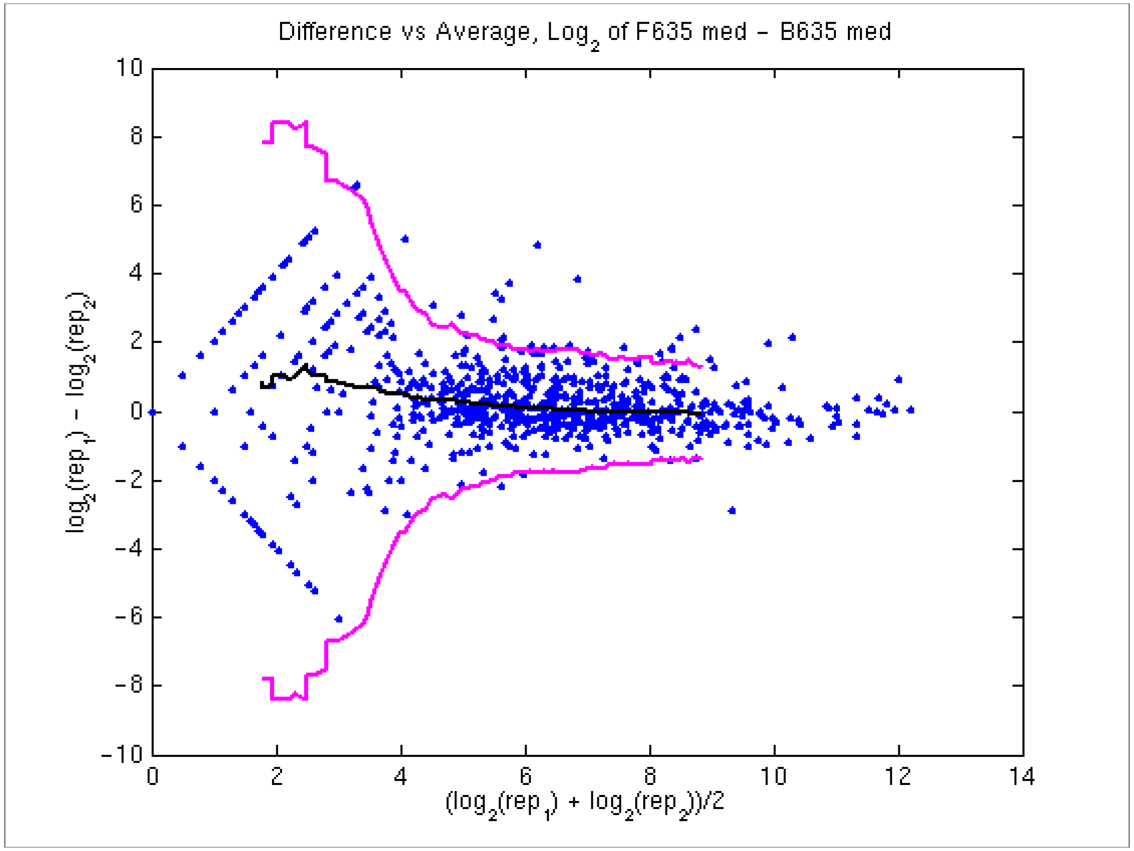
Rep 1



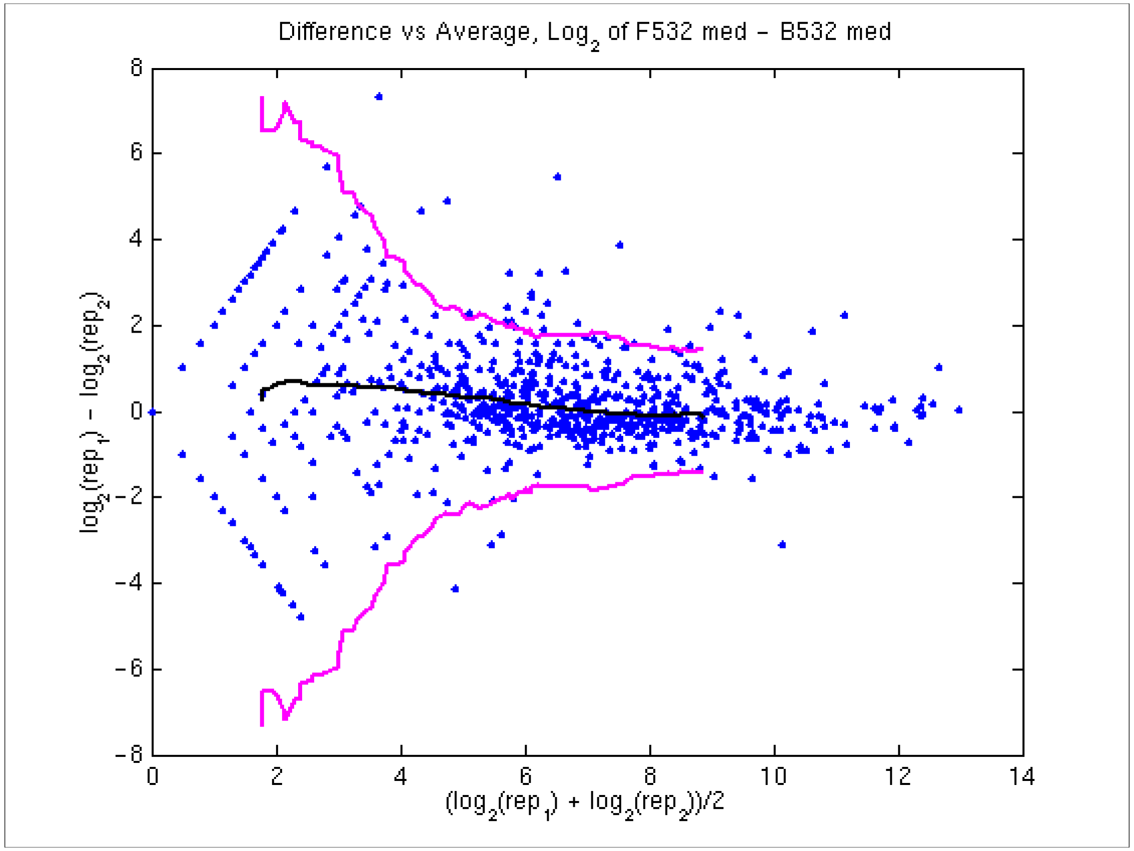
Rep 2



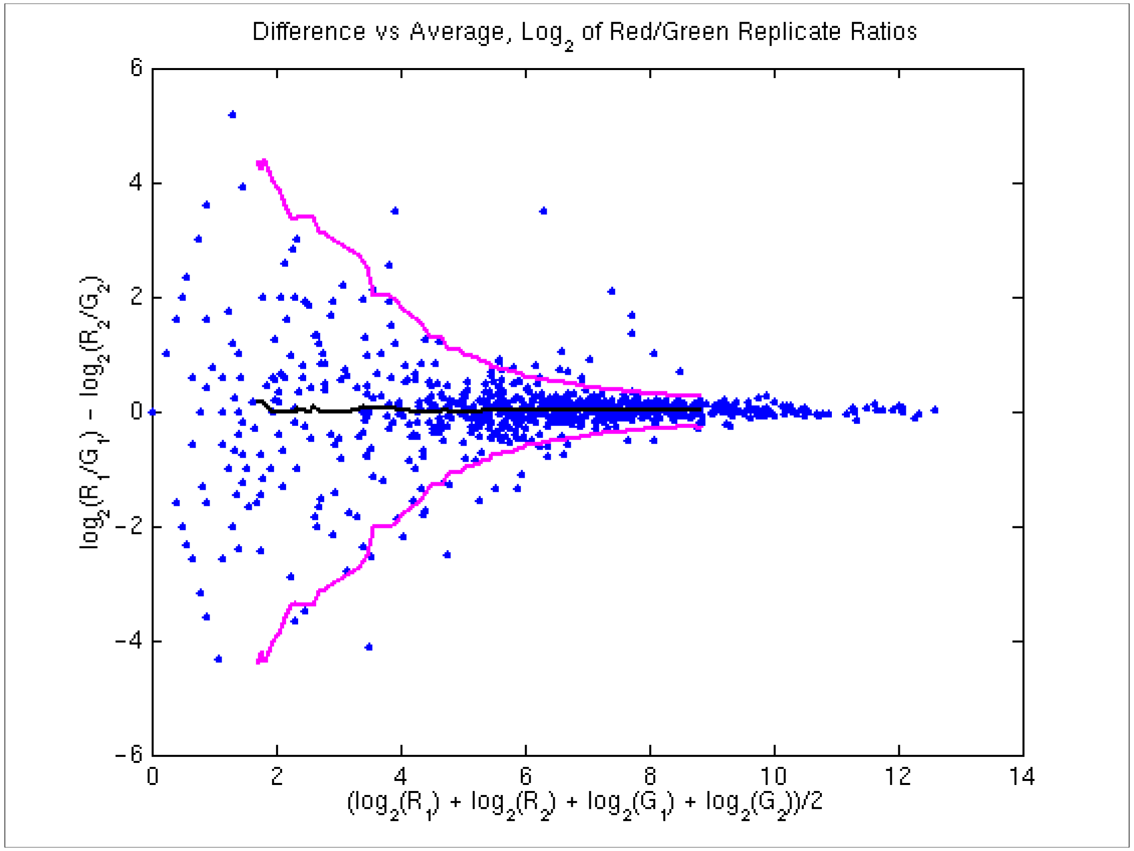
Why Use Log Ratios?: Red Replicates



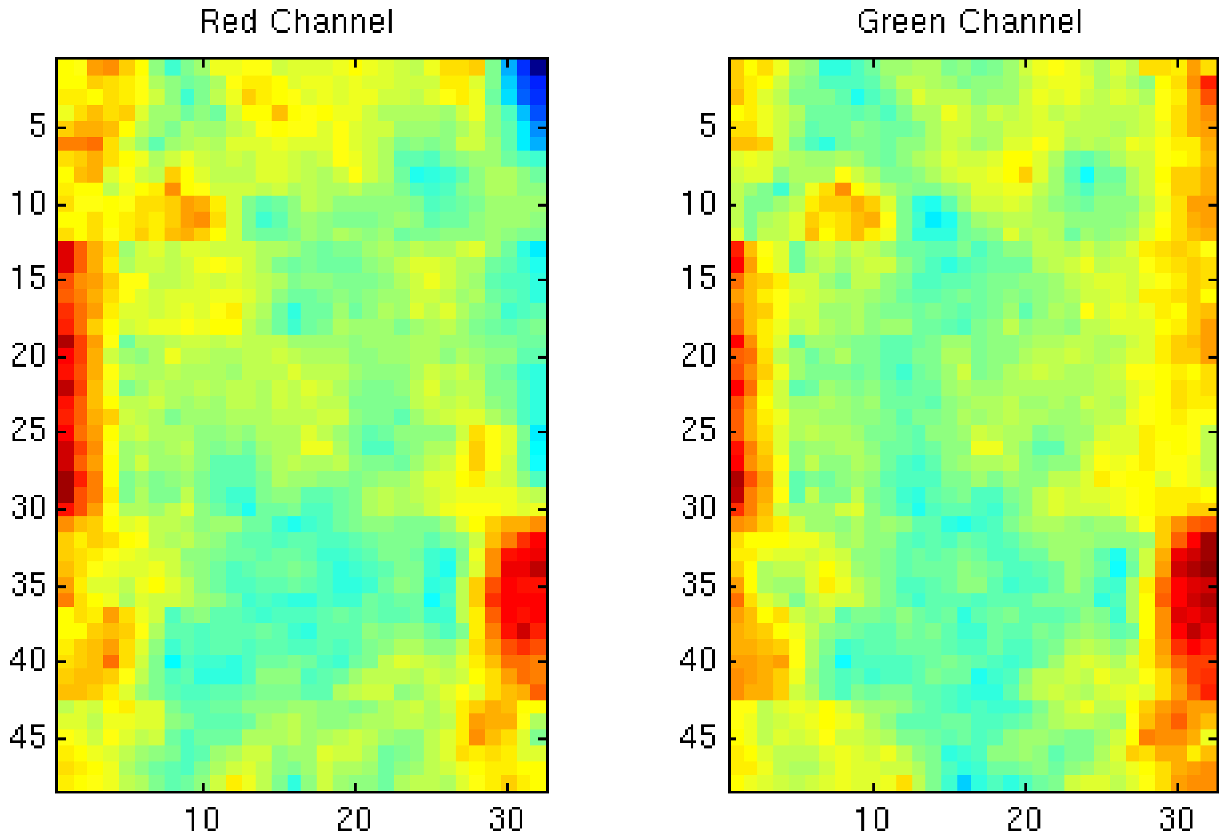
Why Use Log Ratios?: Green Replicates



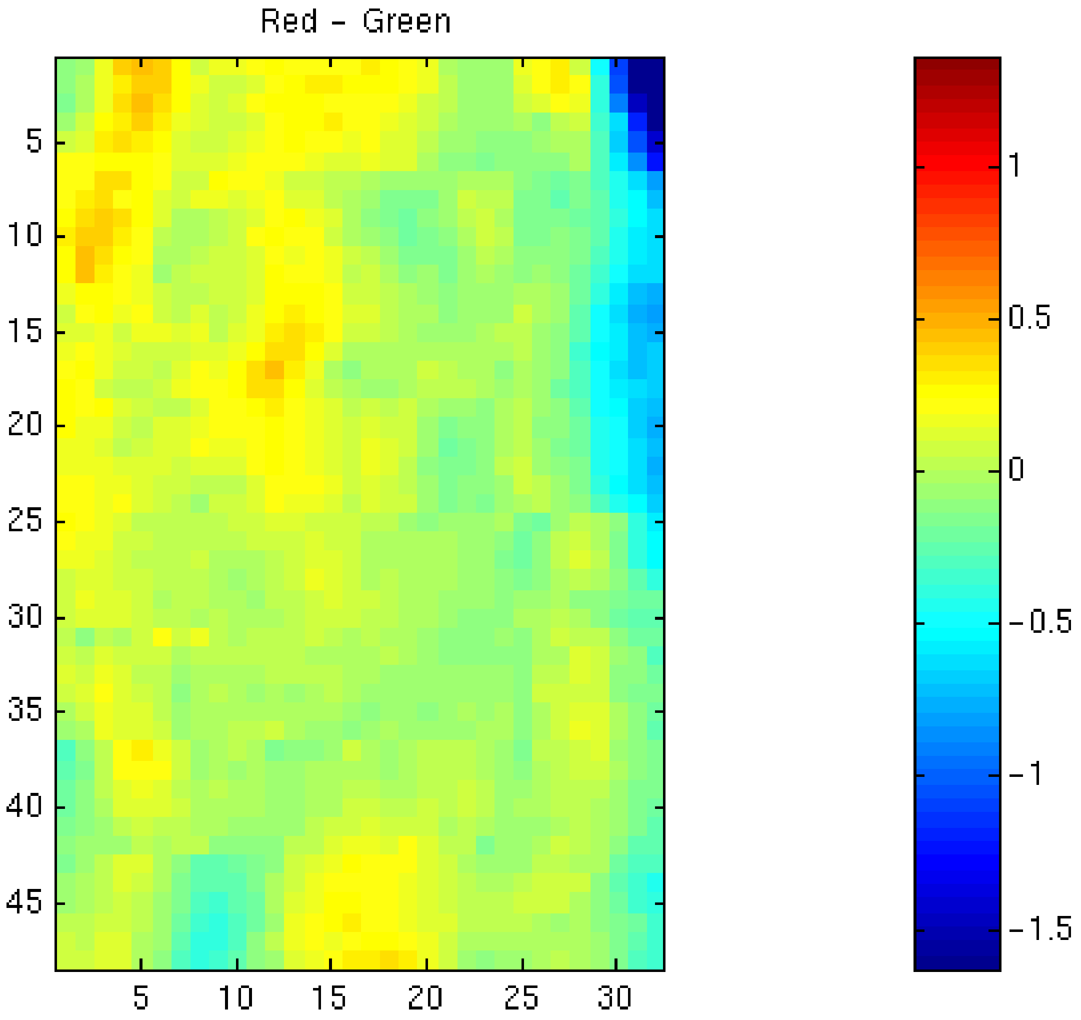
Why Use Log Ratios?: Ratio Replicates



Why does this work? Channels



Why does this work? Ratios


Red - Green