GS01 0163
Analysis of Microarray Data

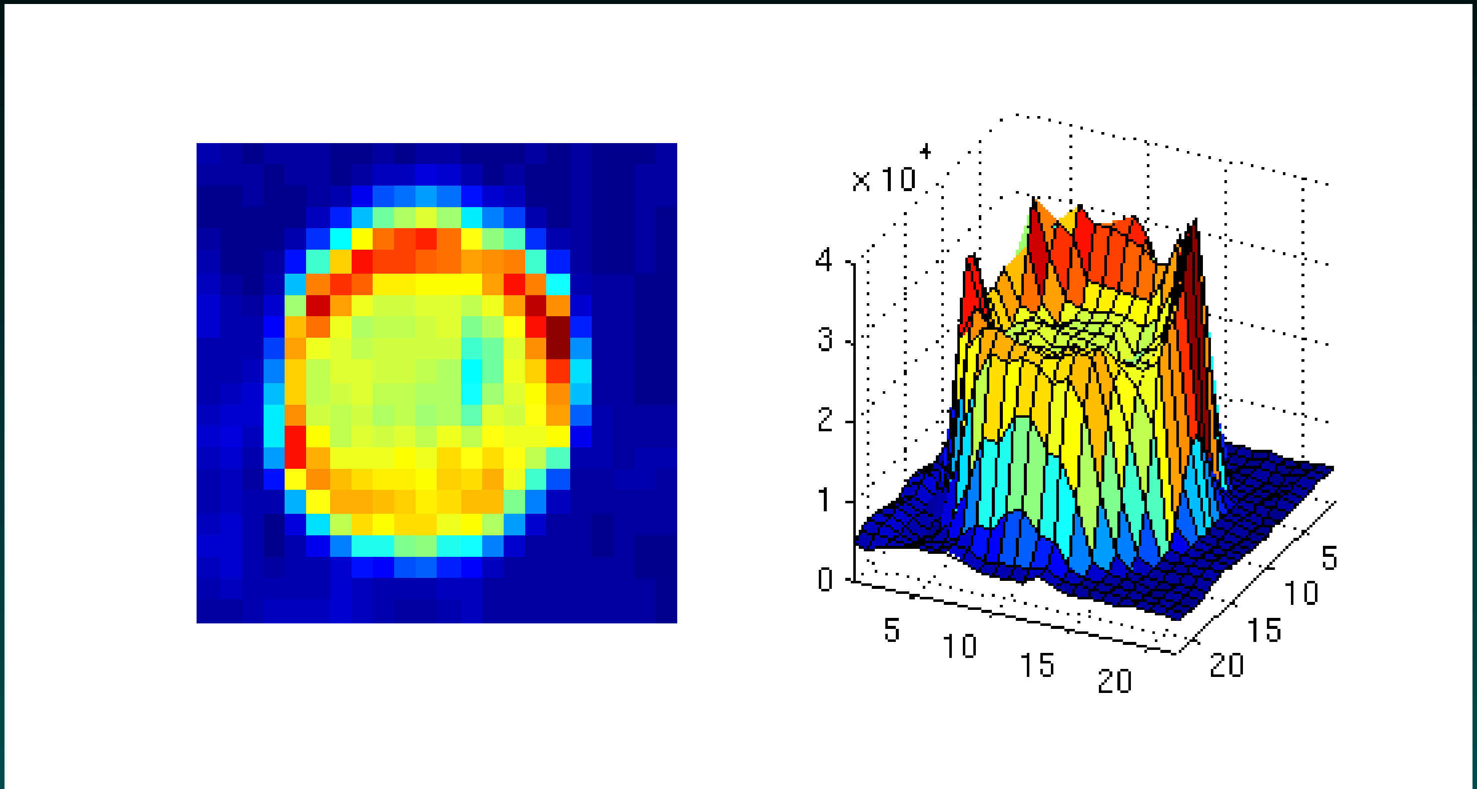
Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

7 September 2004

Lecture 3: Quantifying cDNA 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

The Goal: Understanding 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. . .

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.

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 ($2^{16} - 1$).

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.

Reading Binary Files

At some point, we will assemble a readSimpleTIFF script for R.

In the meantime, we've read some images in Matlab, and exported binary files; there is the `readBin()` function.

For HWK 2 (and here), we'll be looking at matched red and green image files from a single array, and looking at where the quantified numbers come from.

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?

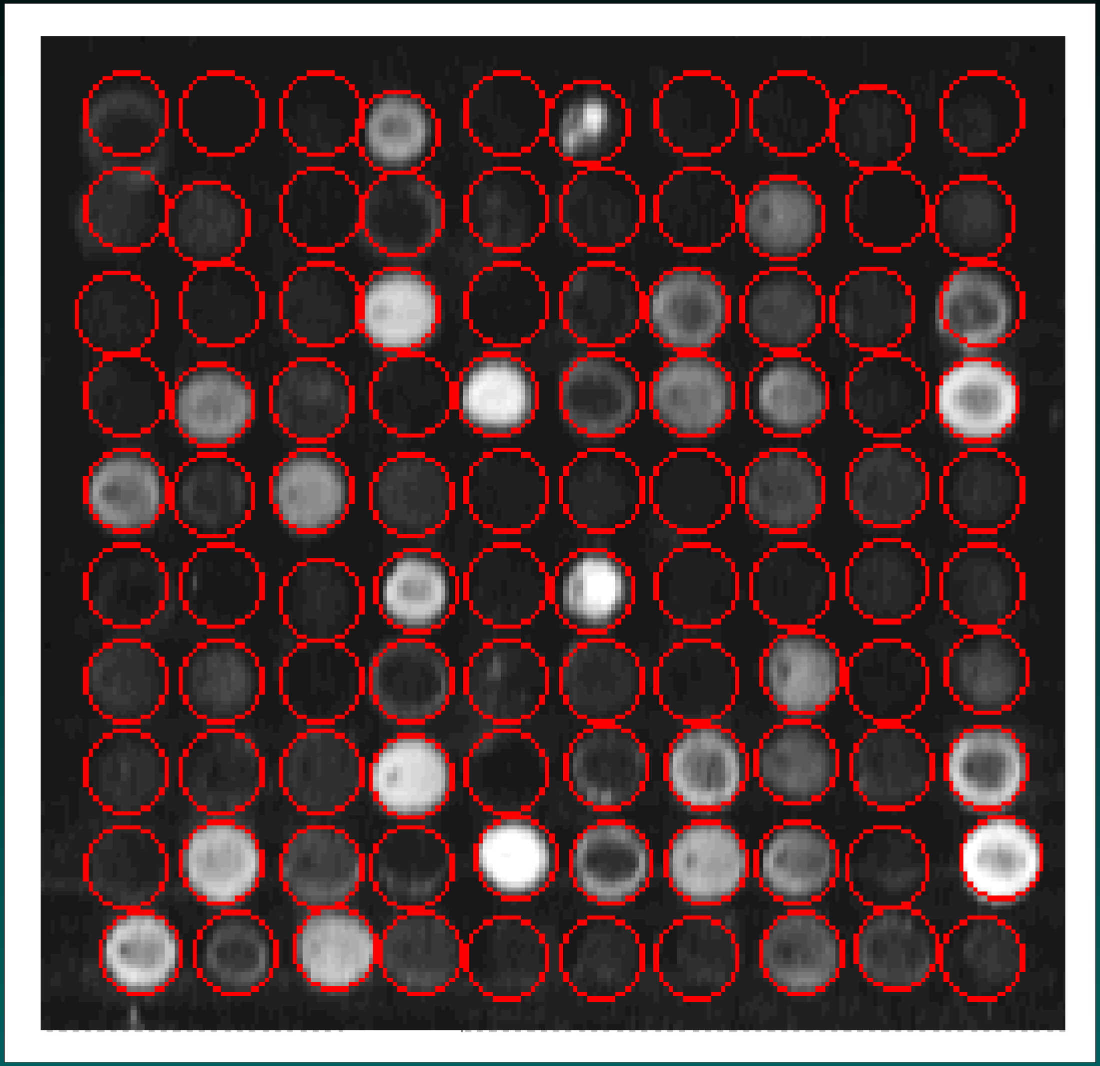
Quantification involves...

registration – locating the centers of the spots

segmentation – deciding which pixels around the center actually belong to the spot

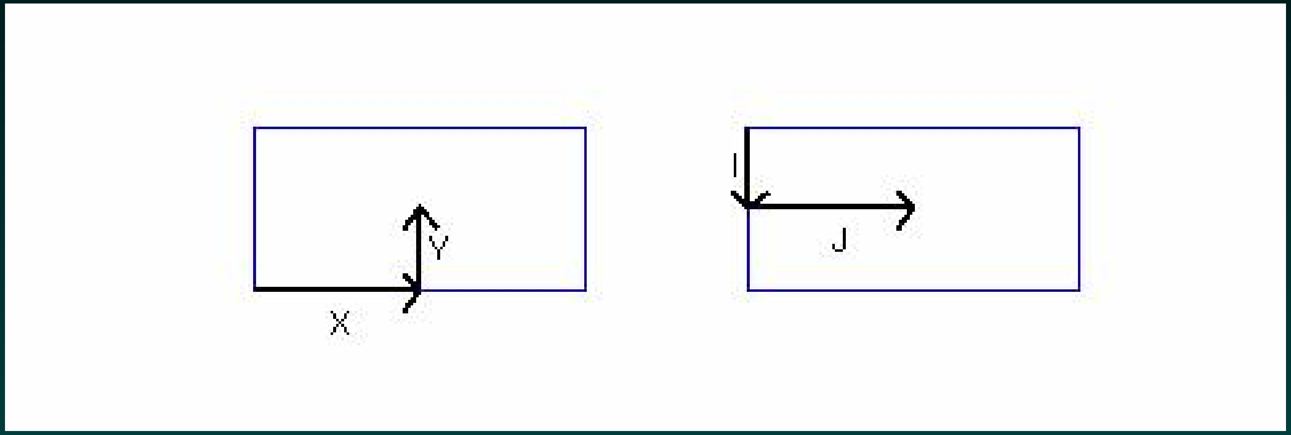
quantification – summarizing the numerical pixel values in the spot (foreground) and around it (background).

a picture of quantification



Position

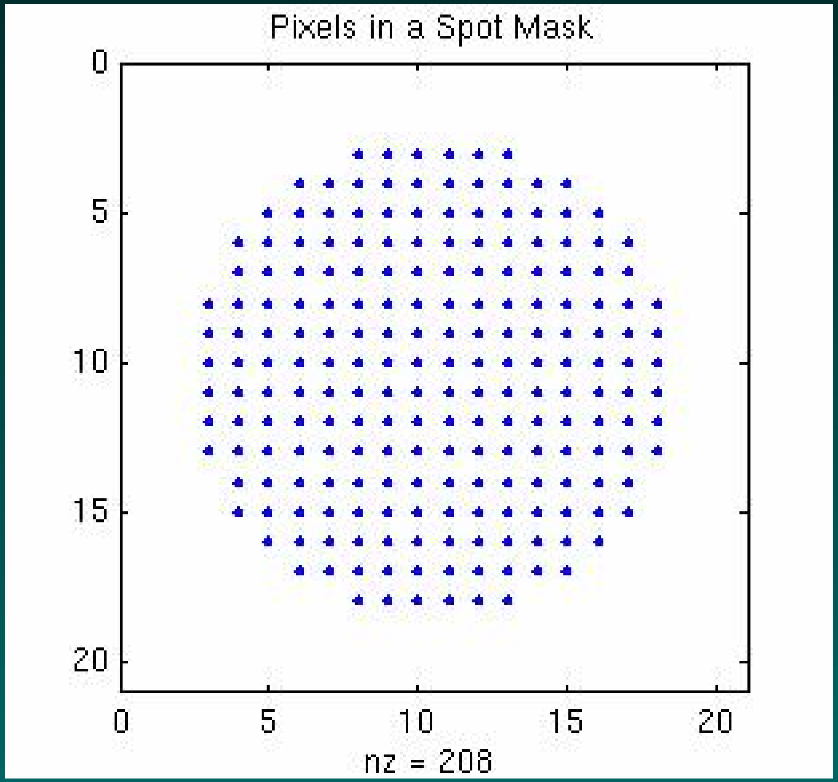
Pos X - mm : 11.278



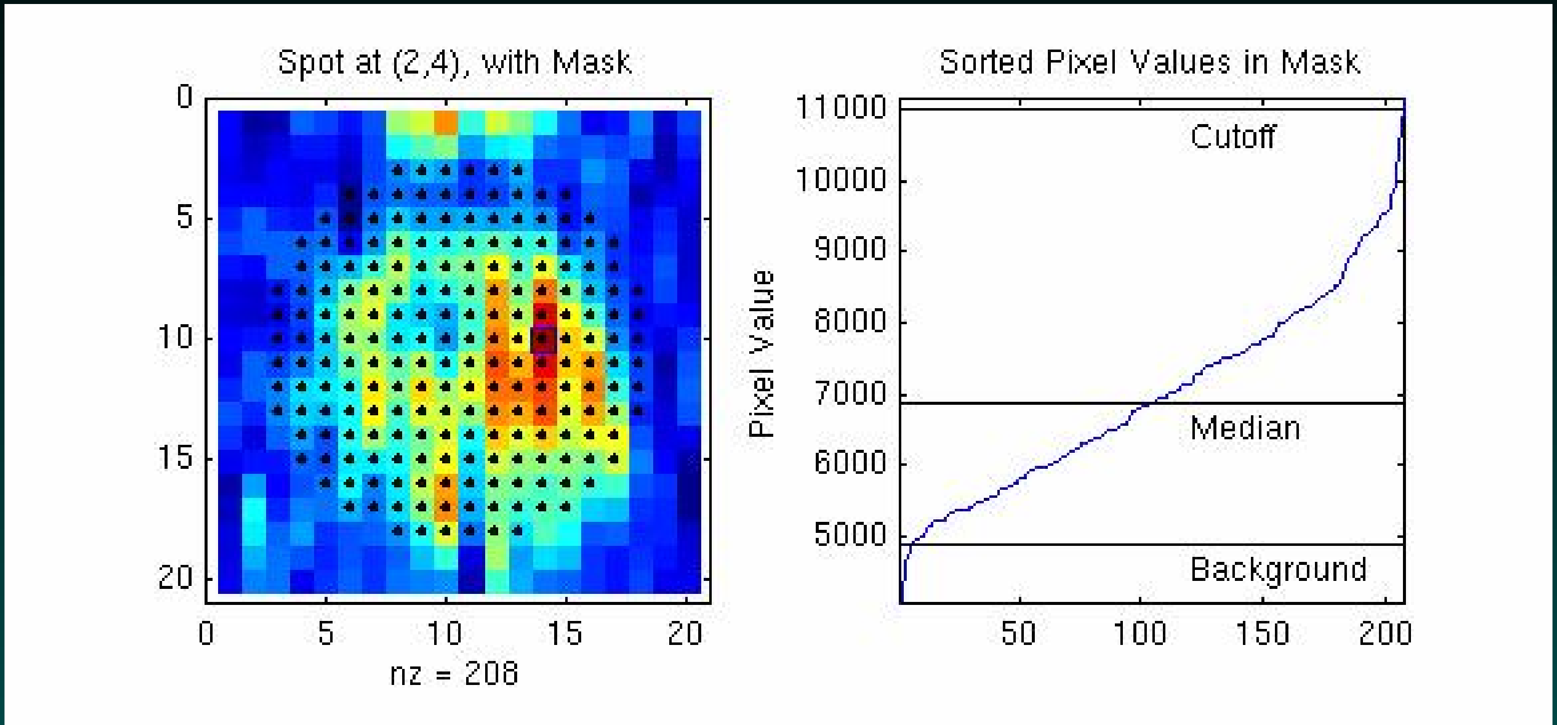
specifying the center of a spot in terms of the slide (mm, from lower left) and image (row and column, from upper left).

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?



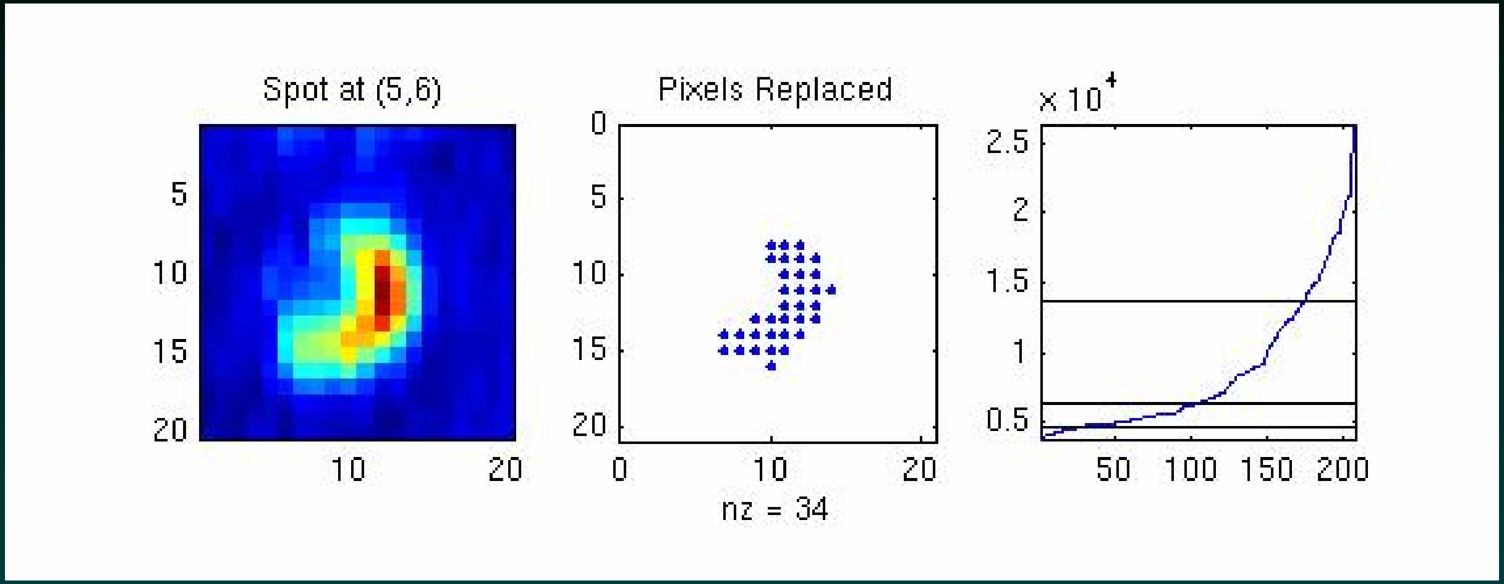
Not quite a perfect fit, but ok.

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.

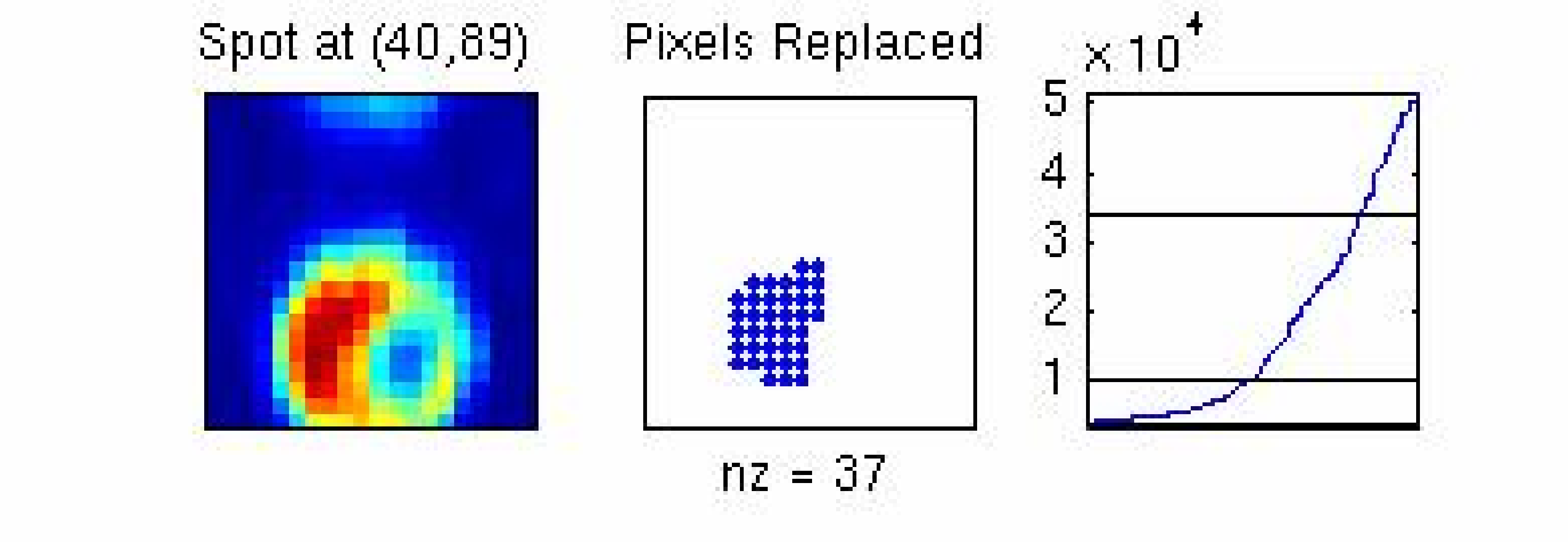
This presumes that the spots are (a) even, and (b) of roughly equal size.

Does AR Vol Work?



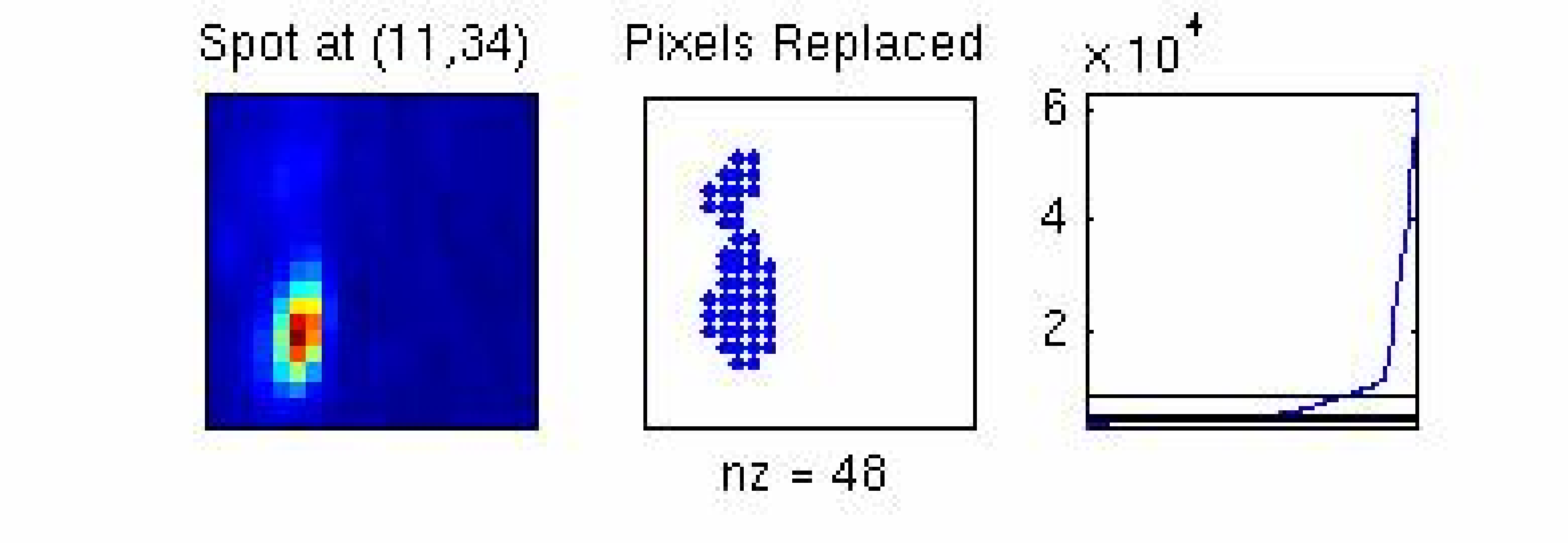
Not always...

Does AR Vol Work?



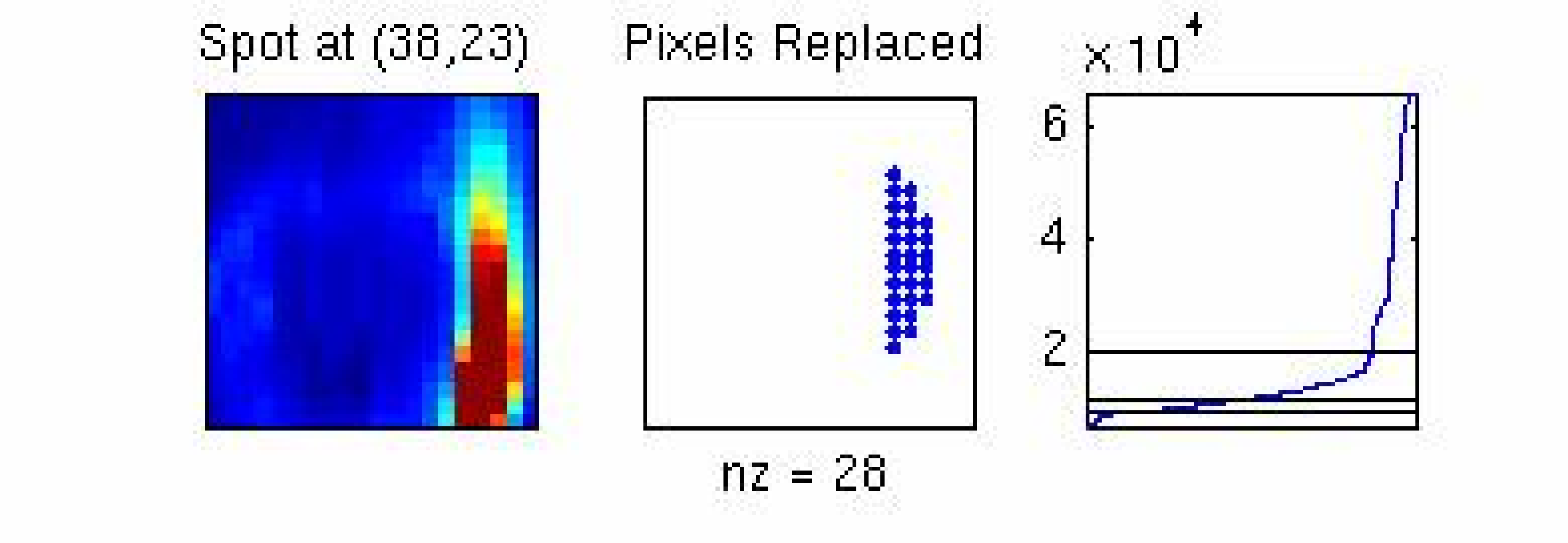
Not always...

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?

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?

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?

What about Background?

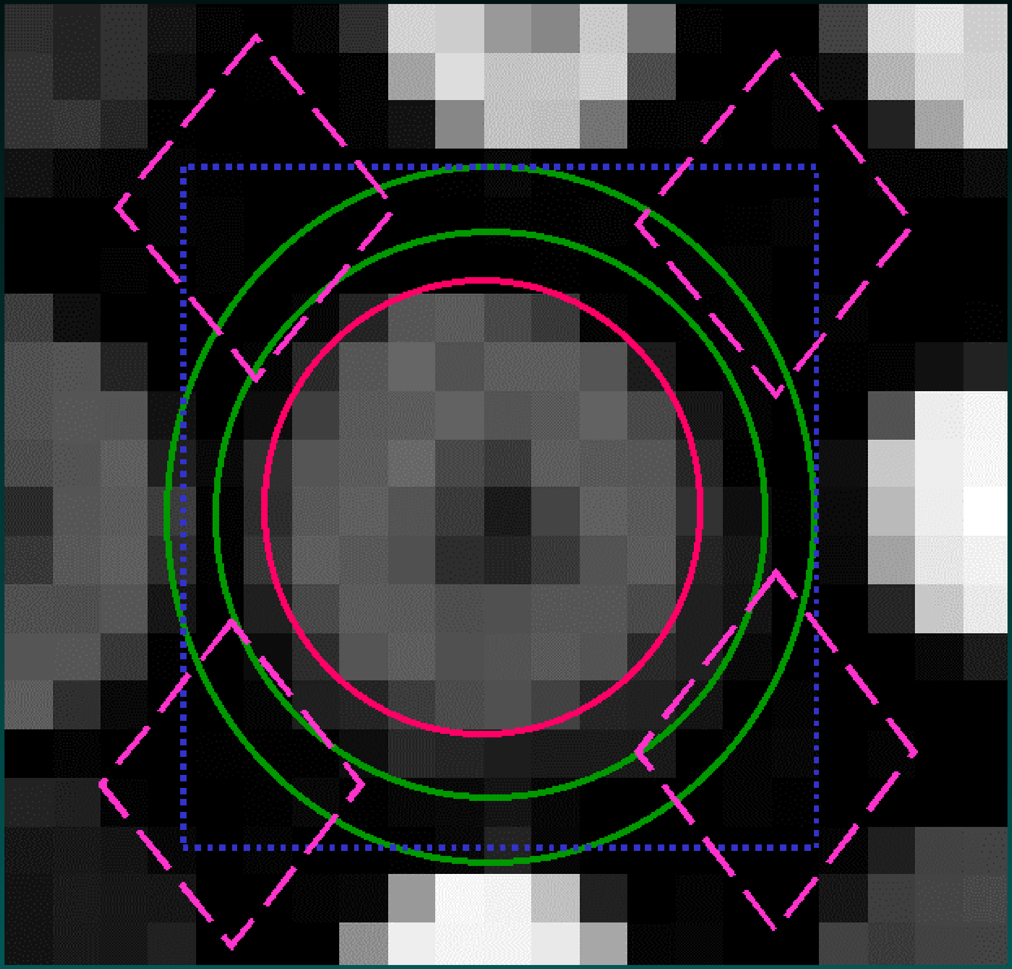
What is it? The pixel intensity in those parts of the image where nothing has been spotted.

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.

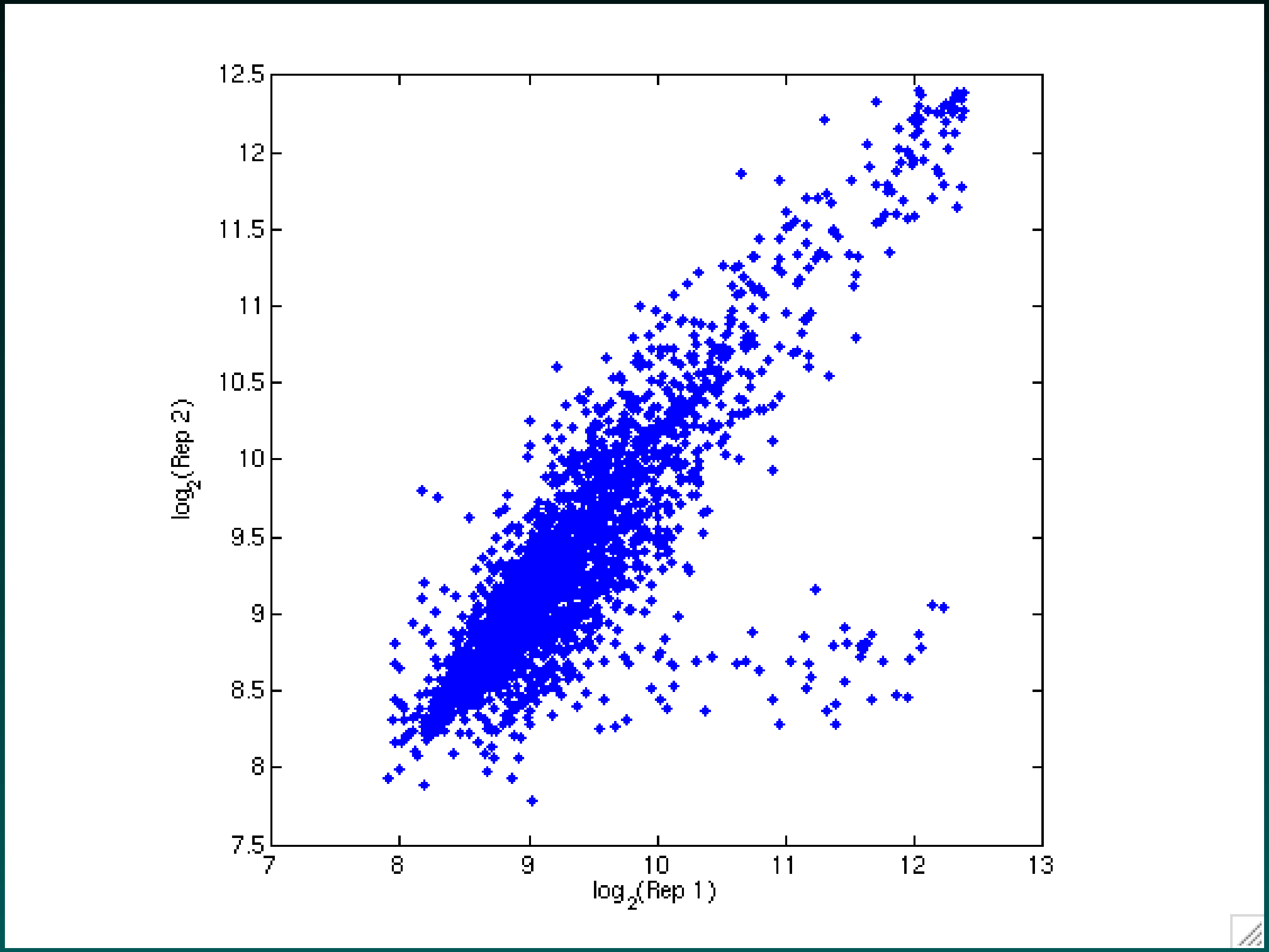
Why log?

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

Why ratios?

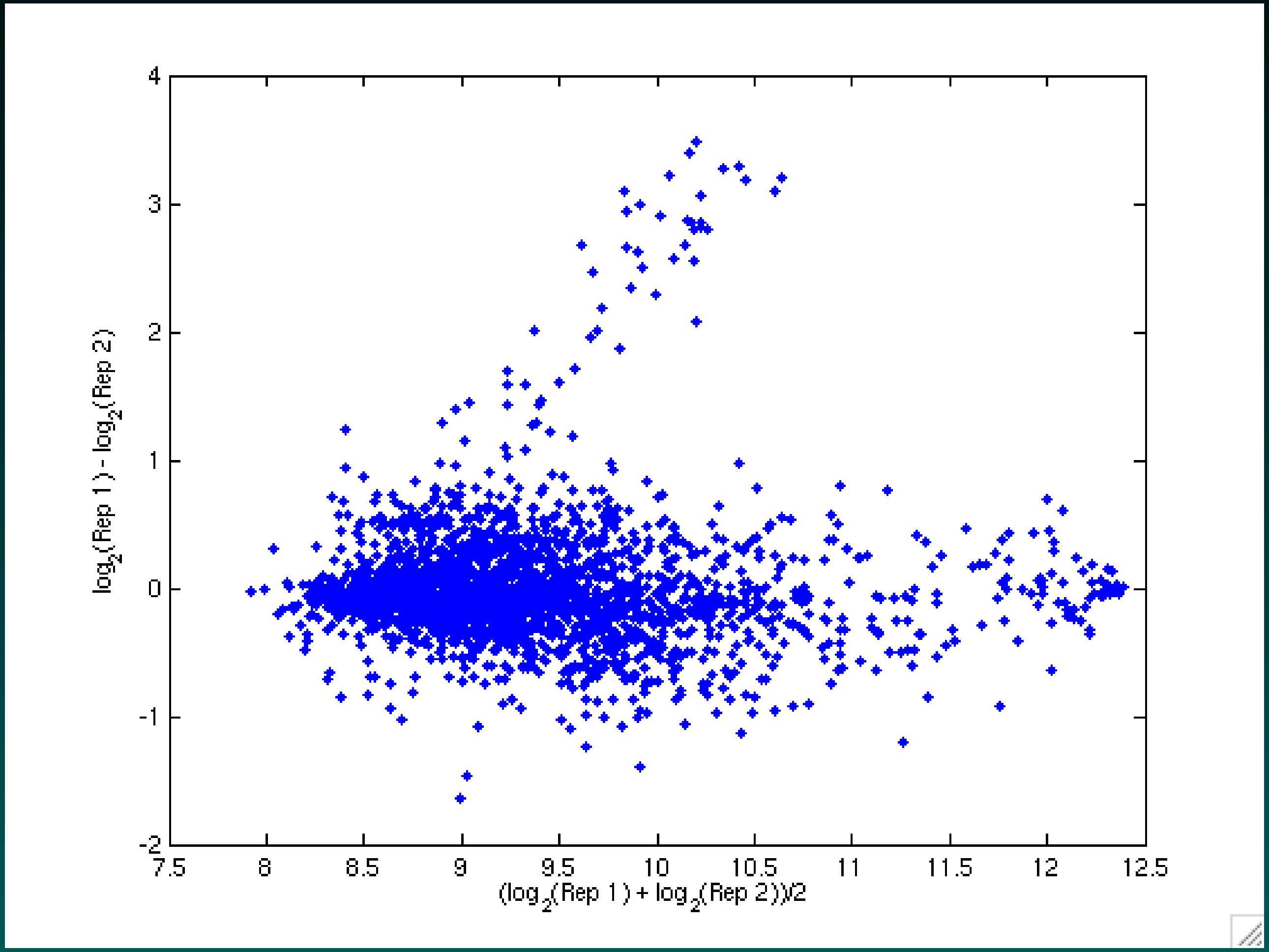
We'll look at that below.

Checking the Data: Replicates



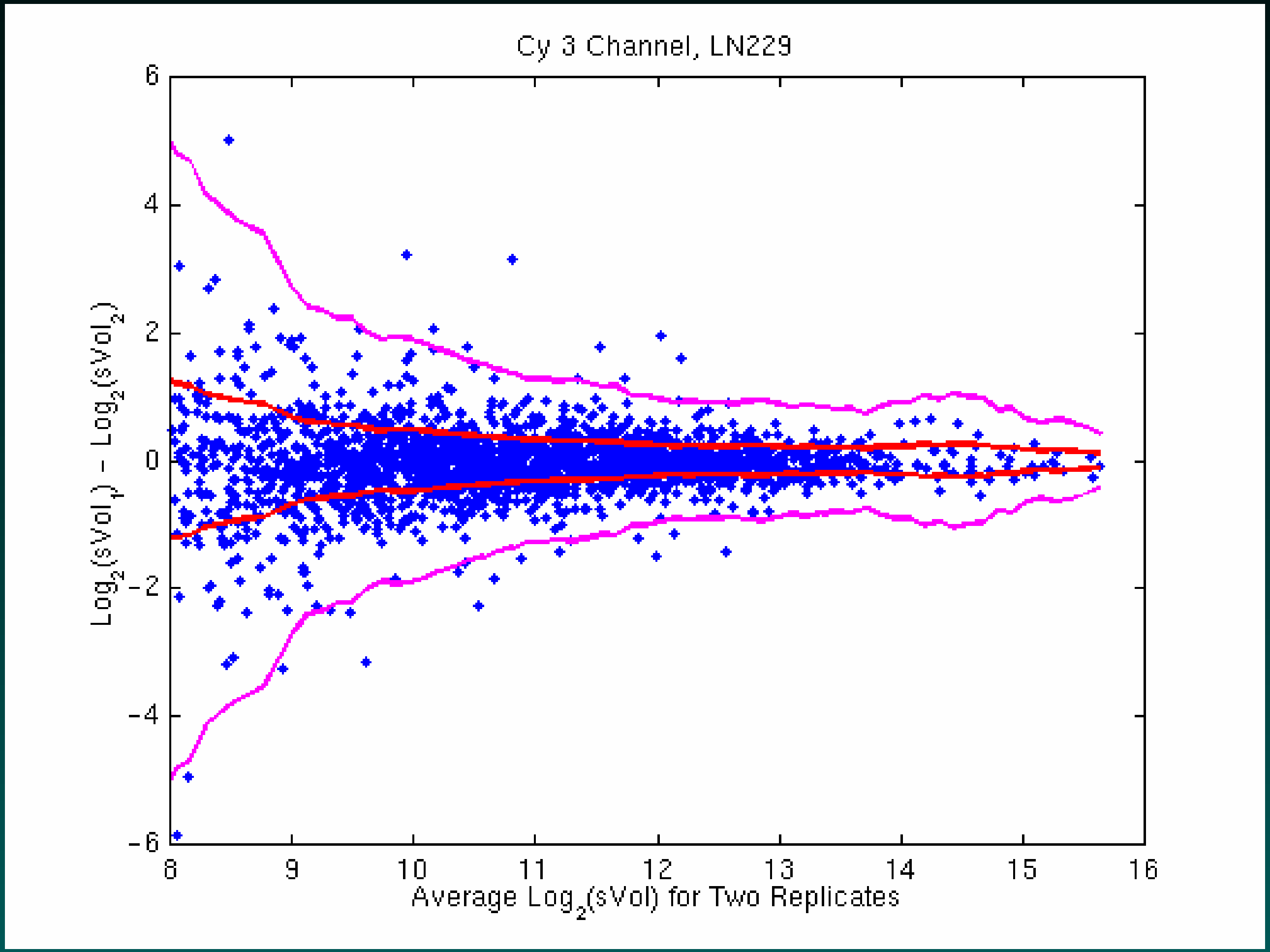
Log scale; do things line up?

A Better View: M-A Plots



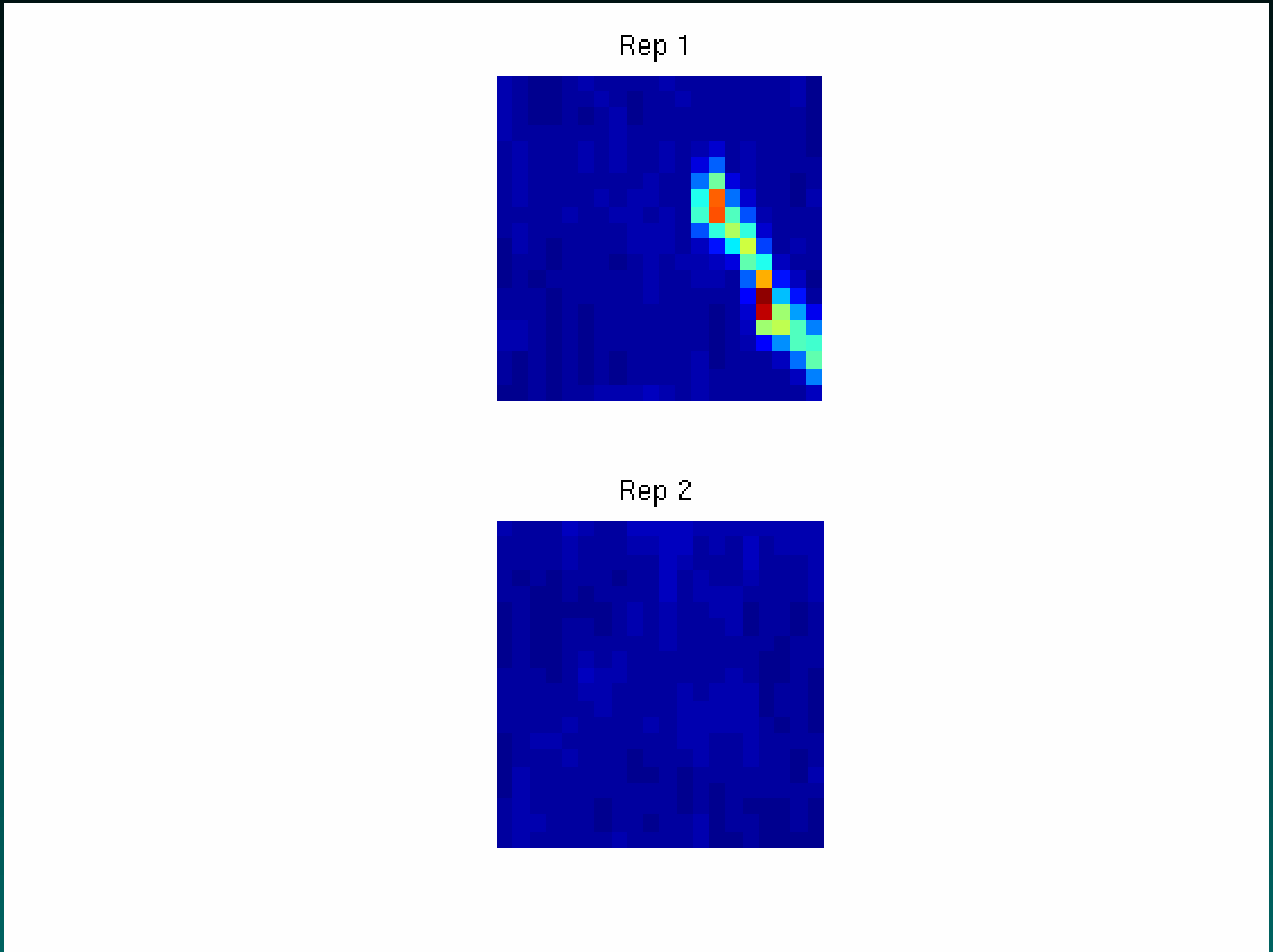
rotate things by 45 degrees so that they’re easier to see.

Subtracting Background, with Replicates

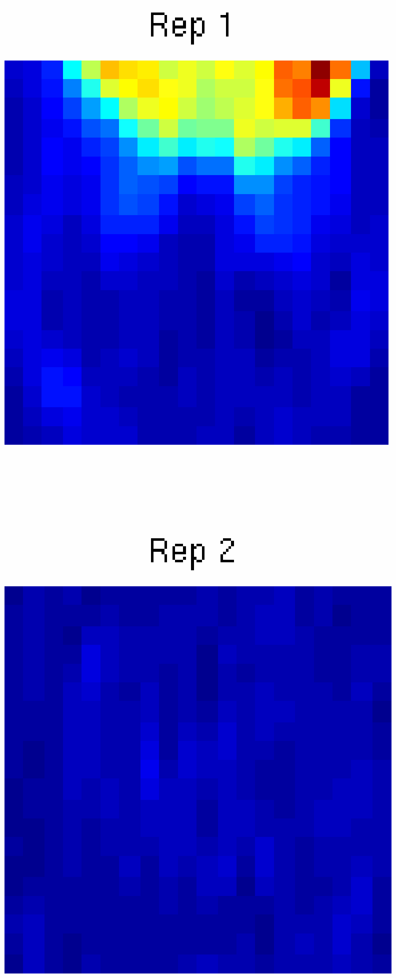


Negative values, thresholding, and variability

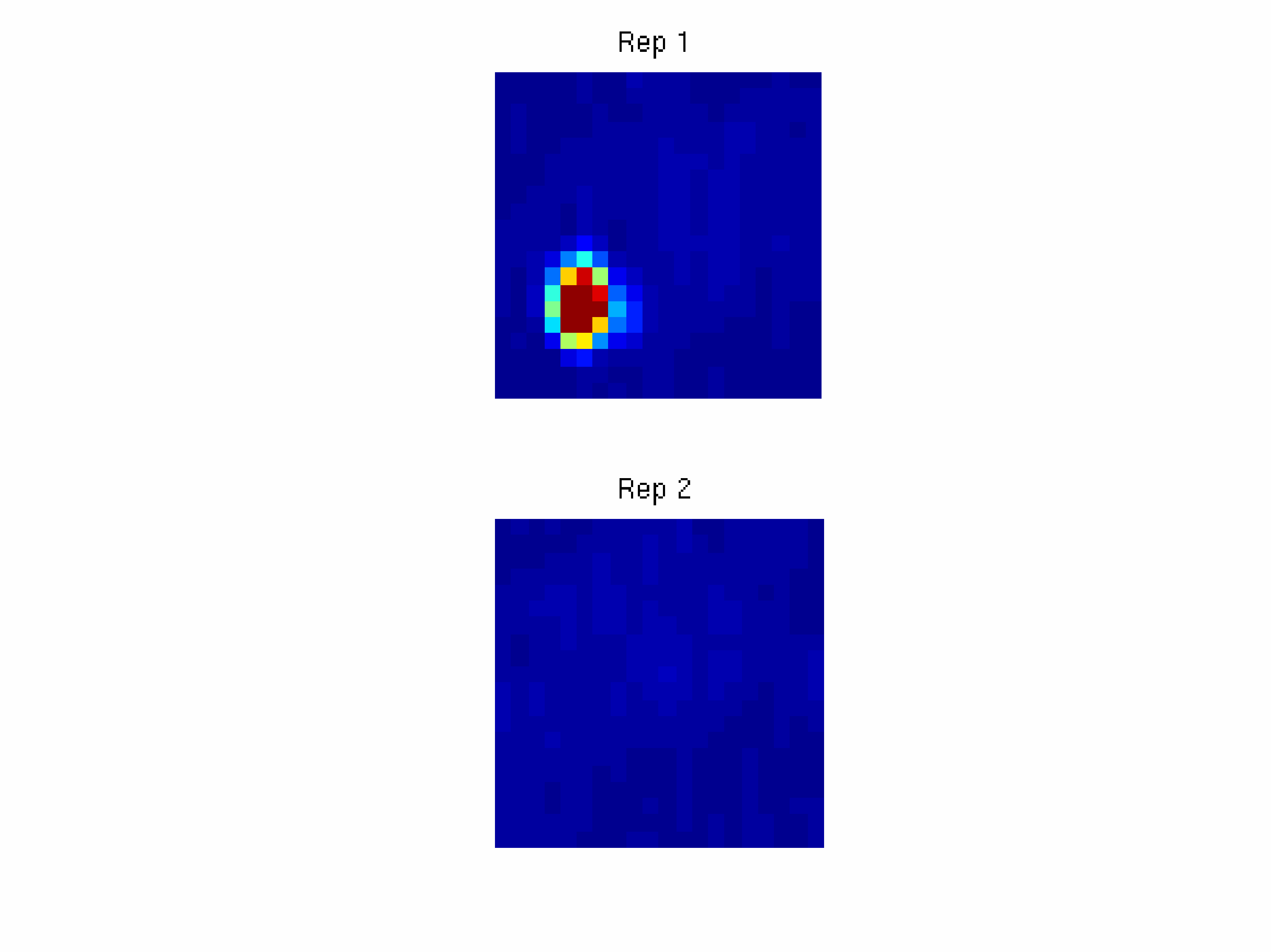
Flagged Spot 1



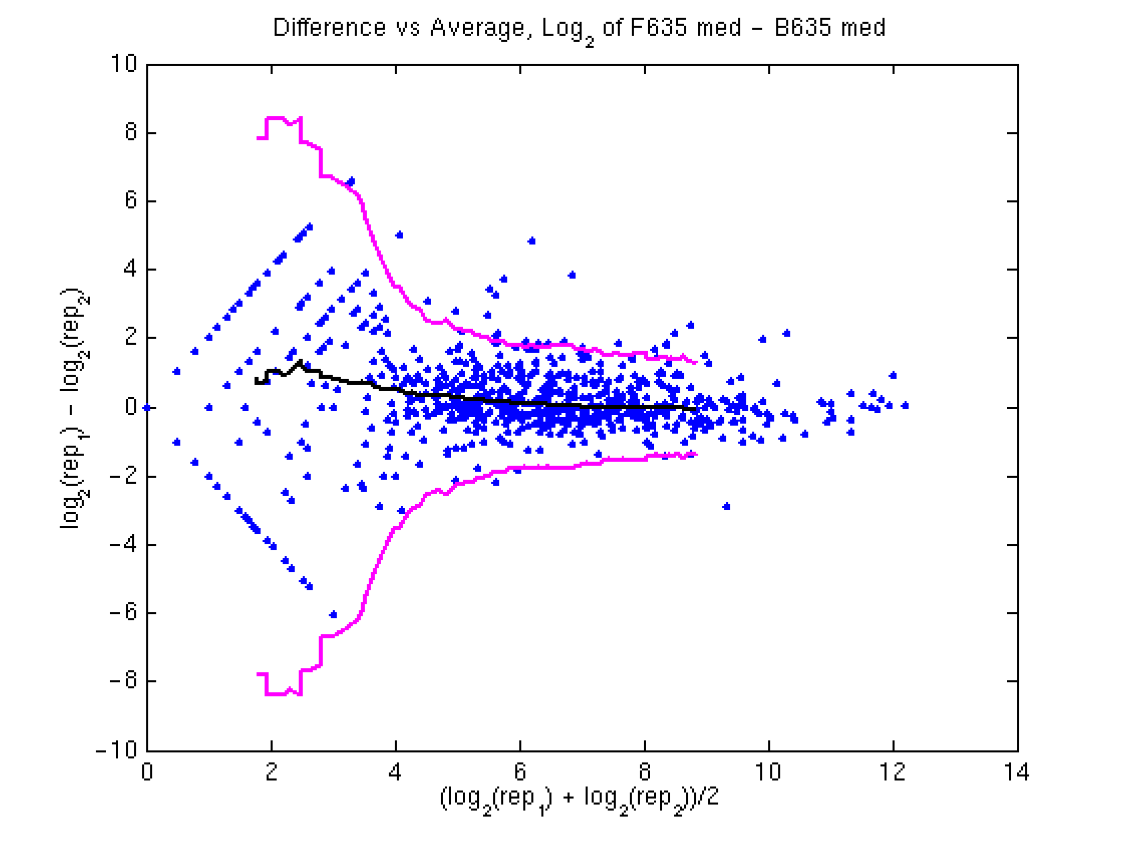
Flagged Spot 2



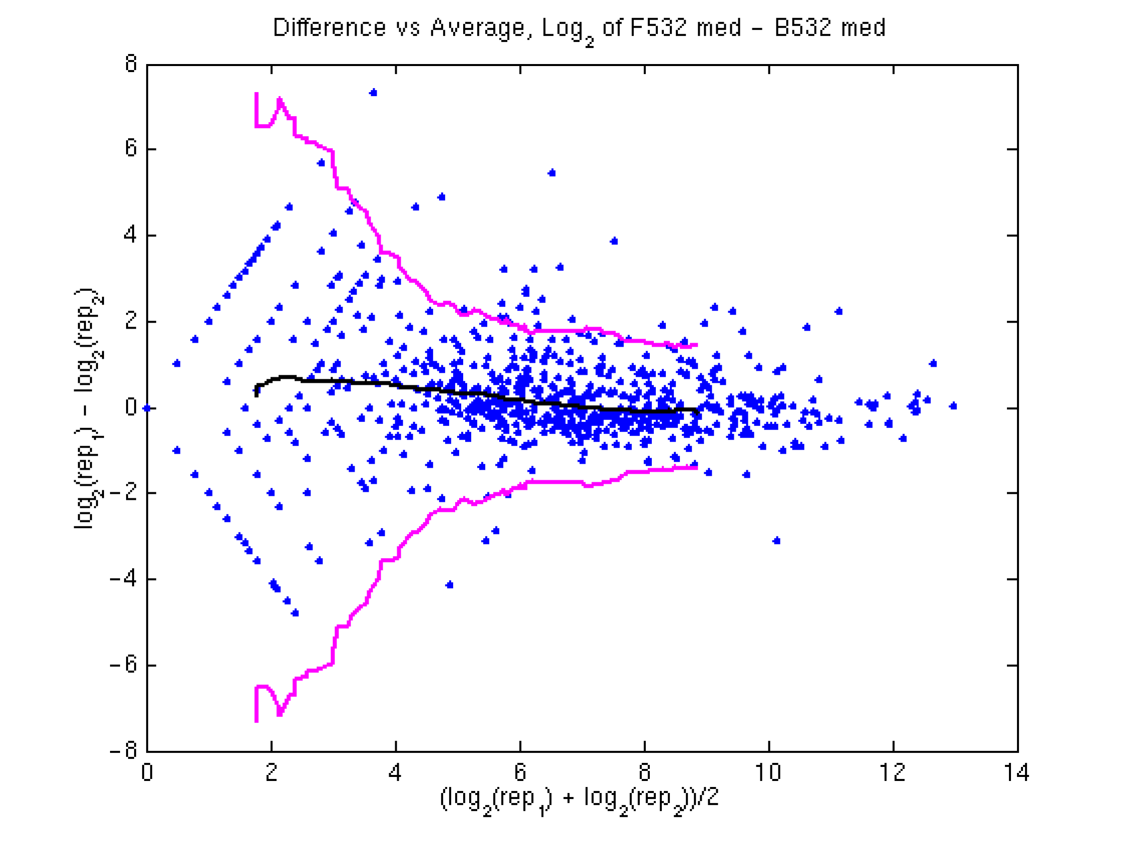
Flagged Spot 3



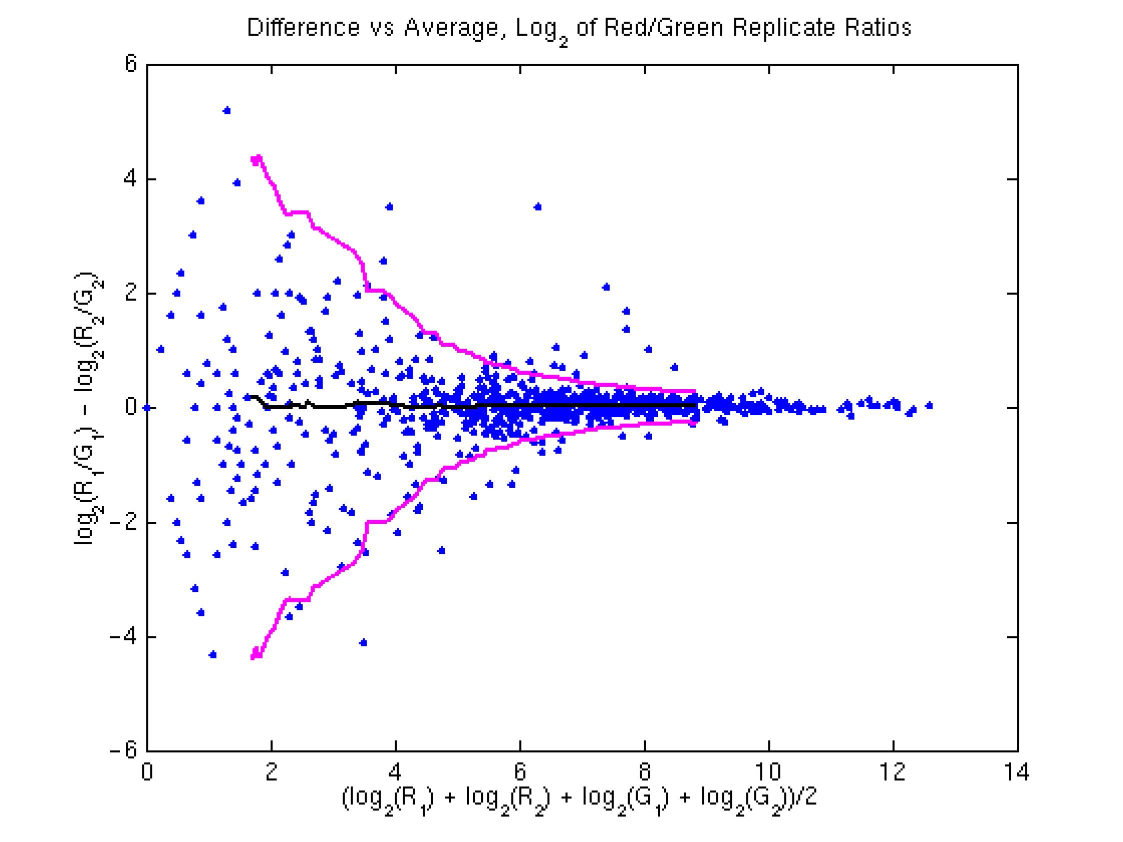
Why Use Log Ratios?: Red Replicates



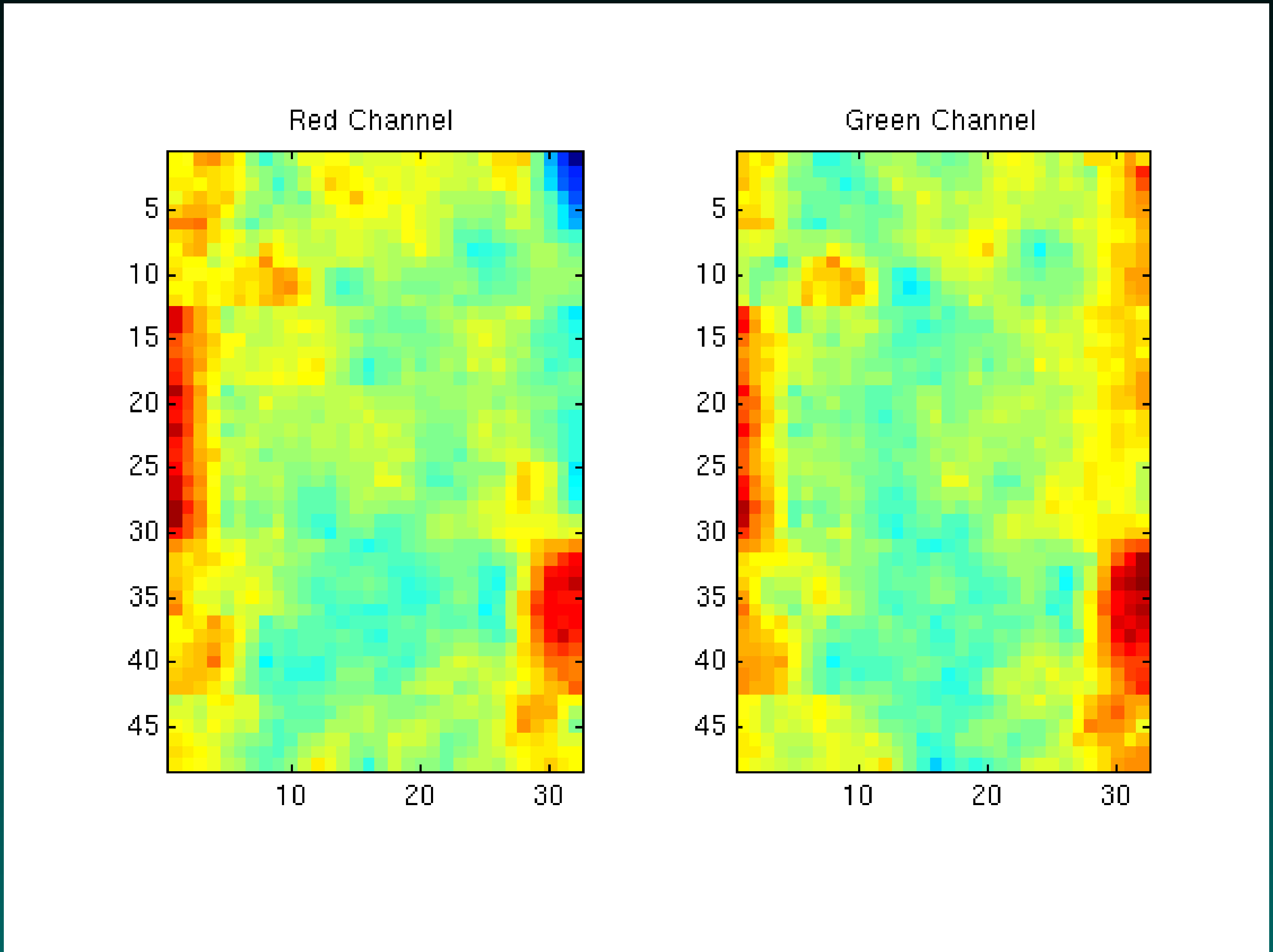
Why Use Log Ratios?: Green Replicates



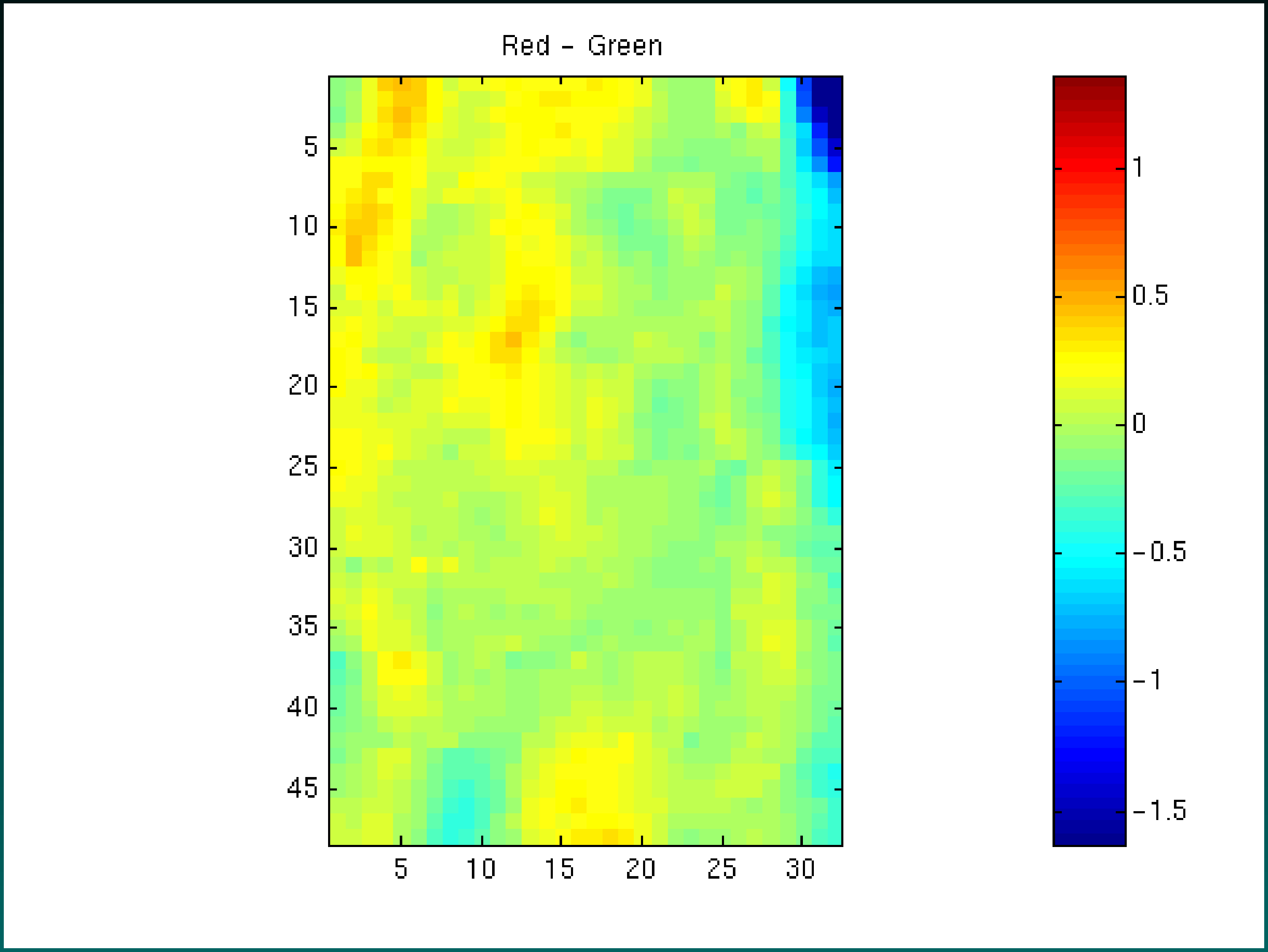
Why Use Log Ratios?: Ratio Replicates



Why does this work? Channels



Why does this work? Ratios


Red - Green