Calling hits on a small molecule microarray

20.109 M1D6 pre-lab

Rob Wilson

SMM Quantification

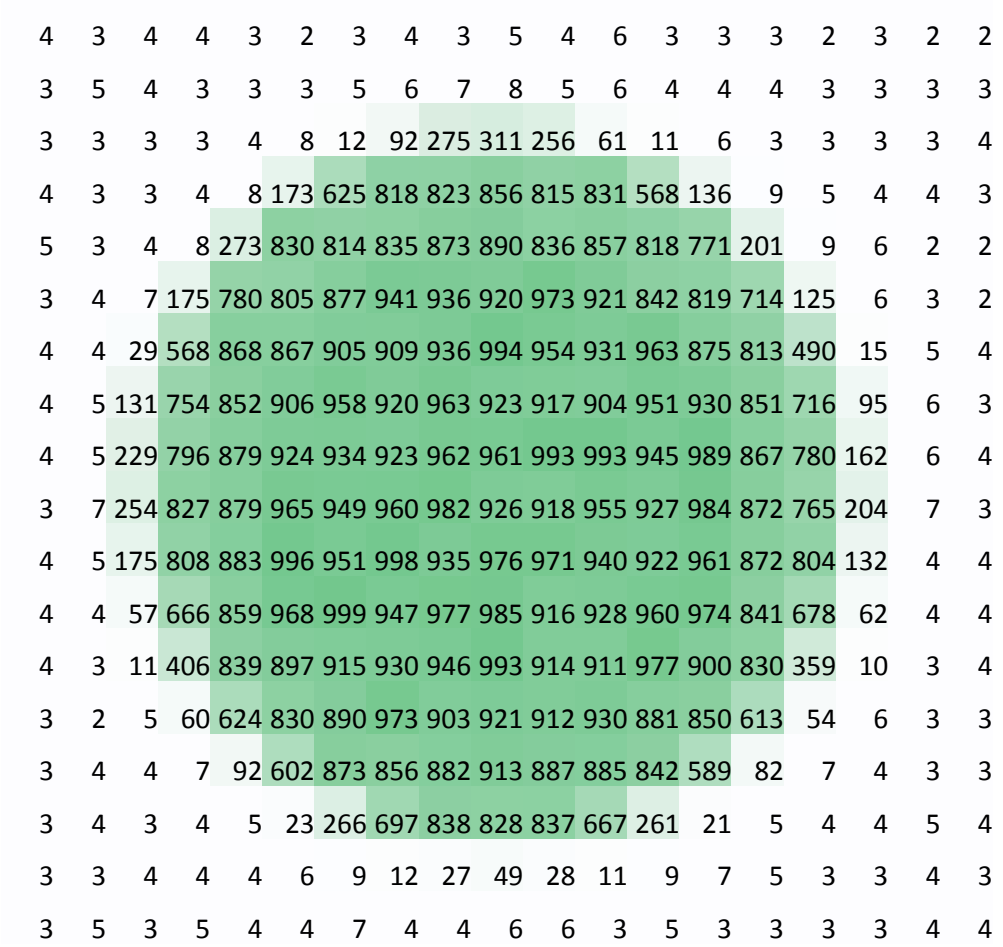
1. Align the GAL file to observed fluorescence on the 532nm channel
2. Quantify fluorescence on the 635nm channel
3. Identify ‘hits’ with improbably high fluorescence
4. Identify compounds which repeatedly hit
5. Analyze top hits for chemical patterns

Images are arrays of numbers

Each pixel is a 16-bit number representing fluorescent intensity

Each slide has two arrays associated with it (one for each excitation wavelength)

These arrays are very large, so we *must* use methods that are computationally efficient

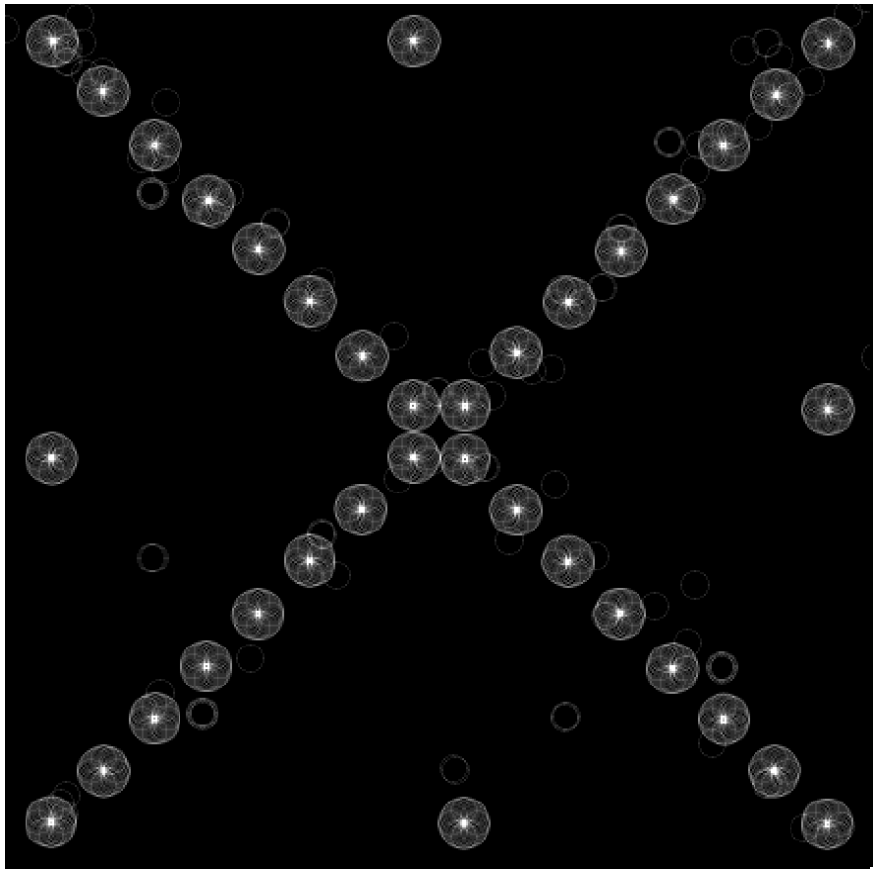


Identifying the “Sentinel” spots

First, find “edge” pixels whose values are greatly different than their neighbors (Canny Edge Detection)

Then, outline each “edge” pixel with a 160µm circle (Hough Transform)

Spot centers are located where lots of circles intersect, and this is extremely insensitive to noise

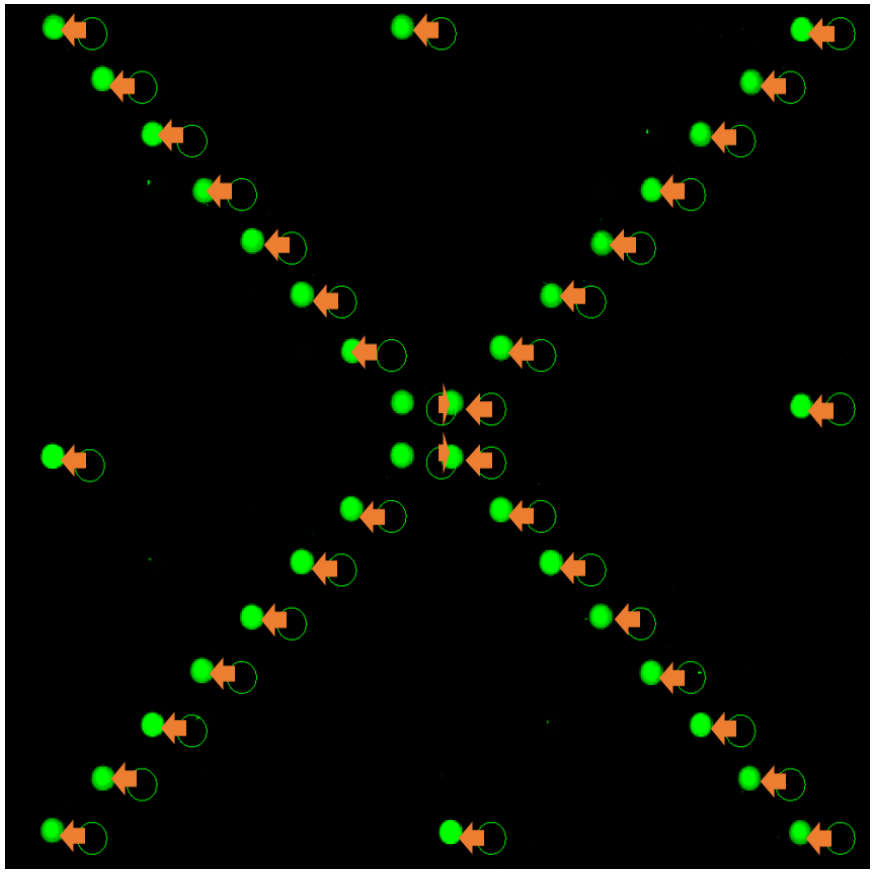


Aligning a GAL file to the Sentinels

1. Pair each predicted spot with the closest observed spot
2. Transform the predicted spots to minimize average distance
3. Repeat until the transformation is negligible

(Iterative Closest Point)

…But you just click the buttons

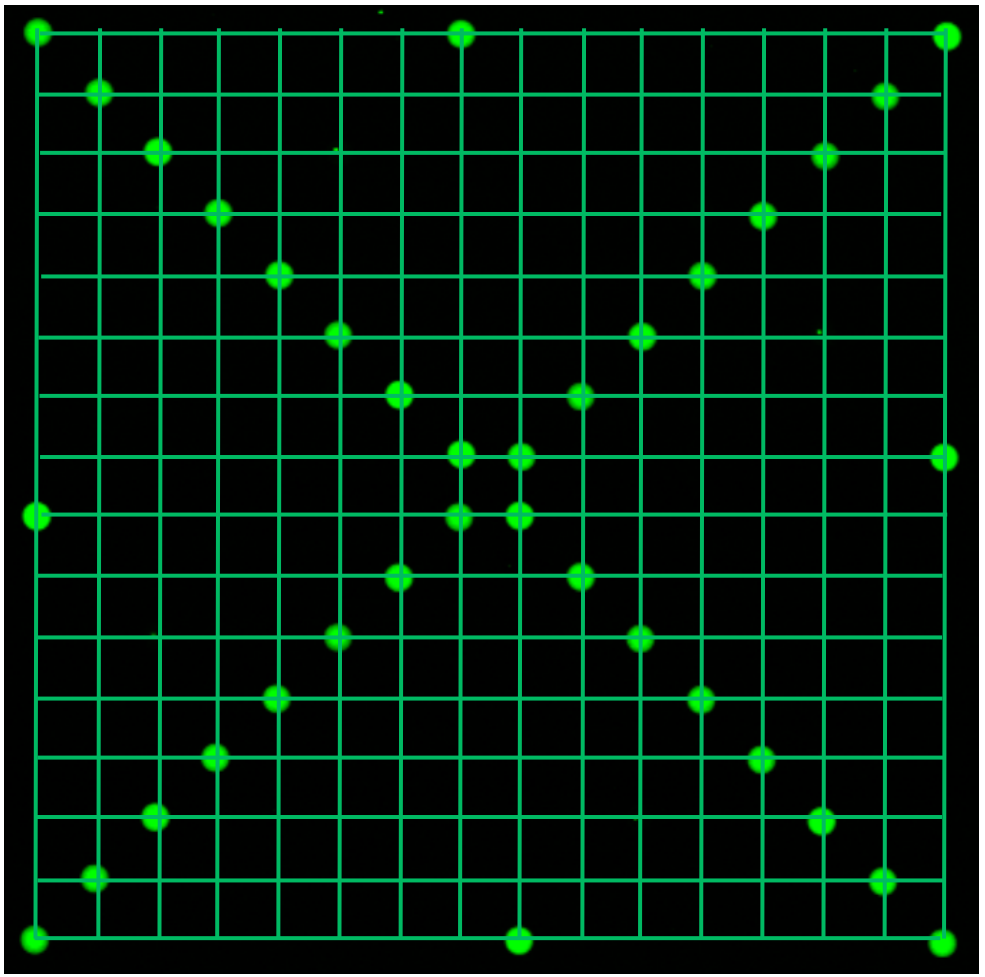


But why so many Sentinels?

Every spot center can be found by intersecting lines between nearby sentinels

Spot center prediction is very precise.

Spot morphology is highly variable.



Quantifying fluorescence intensity

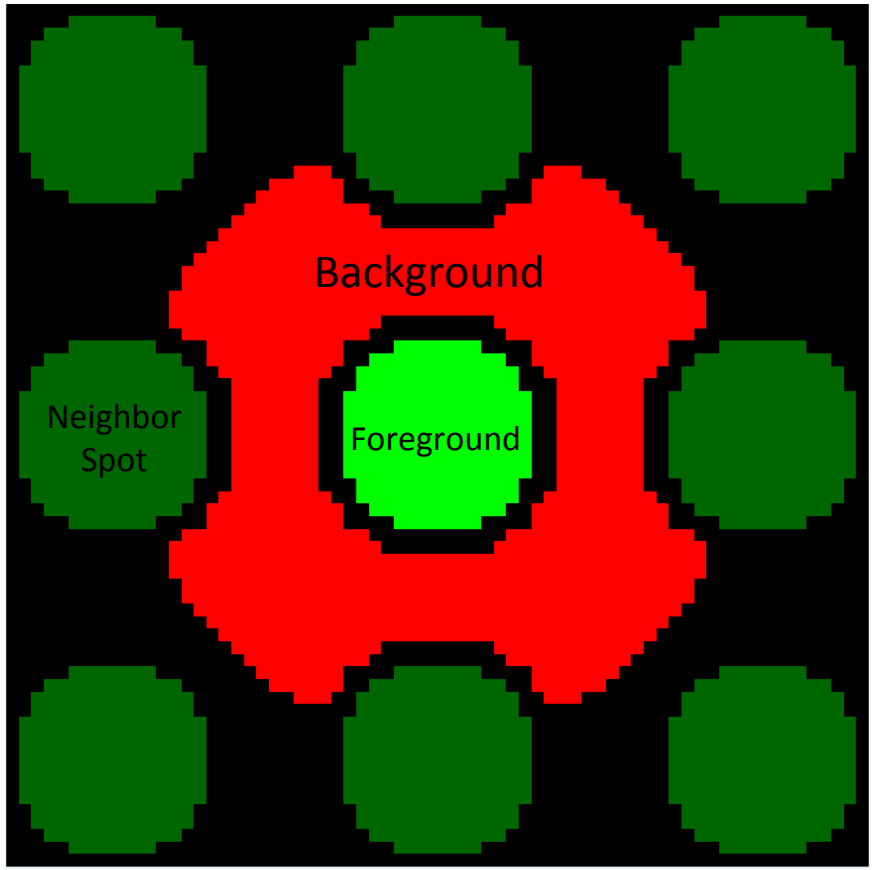
Foreground – Pixels belonging to the printed region, or due to interaction with the compound

Background – Fluorescence not due to interaction with the compound

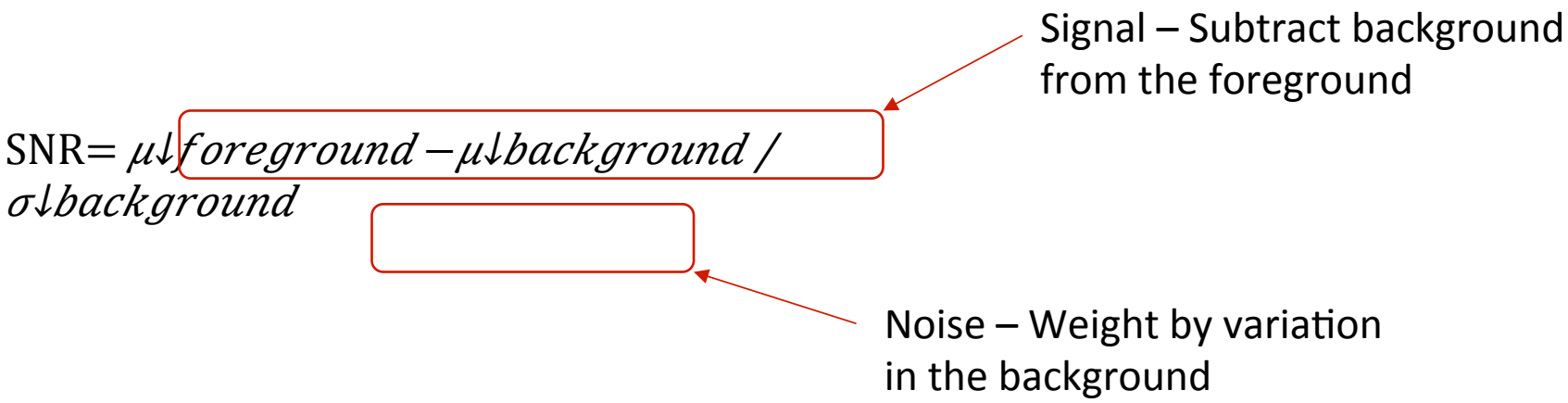
Background fluorescence tends to be residual fluorophore.

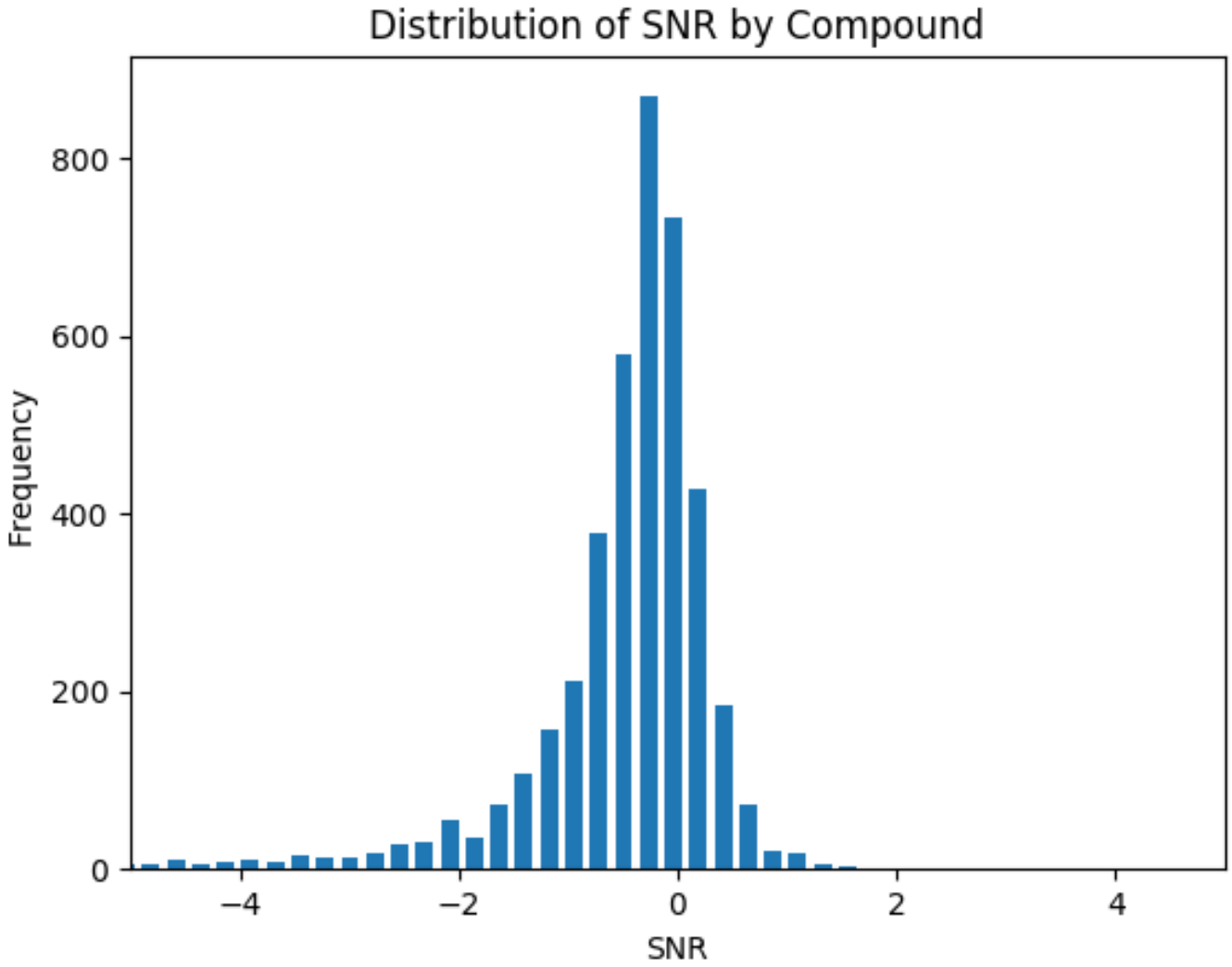
Stringent washing removes low affinity interactions.

We balance these opposing constraints by running test slides before each screen.



Signal-to-Noise Ratio (SNR)





This isn't quite a Normal distribution, but it's pretty close.
The data is skewed, but we can still calculate Robust Z scores.

Z Scores vs. Robust Z Scores

	Z Score	Robust Z Score
Deviation	$x_i - mean(x)$	$x_i - median(x)$
Measure of Distribution	$\sqrt{\sum (x_i - mean(x))^2 / N}$	$median(\lvert x_i - median(x) \rvert)$ (Median Absolute Deviation)
Final Z value	$(x_i - mean(x)) / \sqrt{\sum (x_i - mean(x))^2 / N}$	$(x_i - median(x)) / (median(\lvert x_i - median(x) \rvert) \times 1.48)$

MAD can be used for any distribution; 1.48 is a scale factor for the Normal Distribution
Robust Z scores eliminate the influence of outliers.

Factors influencing hit-calling

How many false positives do we expect?
- We need more hits if we're not confident

How many chemical patterns can we recognize?
- Repeated patterns increase confidence

Are the hits we observe unique to this screen?
- Promiscuous binders are not desirable

Can we get or make more of the compounds?
- Commercial availability and synthetic tractability limit development

How many can we afford to advance?
- Secondary assays are expensive, so we can't advance more than we can afford

Evaluating an SMM

Is the background constant or noisy?

Can the positive controls be easily recognized?

Are there any manufacturing defects? Handling defects?

When looking at hits, where are the printing sites?

Do you trust the data?