

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

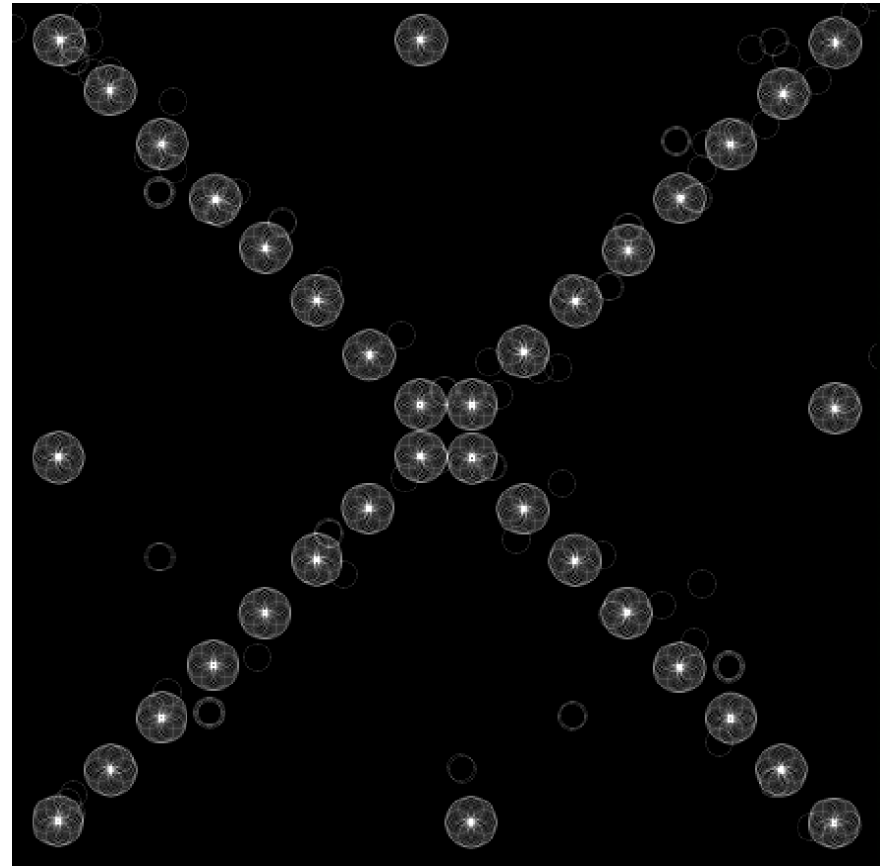
4	3	4	4	3	2	3	4	3	5	4	6	3	3	3	2	3	2	2
3	5	4	3	3	3	5	6	7	8	5	6	4	4	4	3	3	3	3
3	3	3	3	4	8	12	92	275	311	256	61	11	6	3	3	3	3	4
4	3	3	4	8	173	625	818	823	856	815	831	568	136	9	5	4	4	3
5	3	4	8	273	830	814	835	873	890	836	857	818	771	201	9	6	2	2
3	4	7	175	780	805	877	941	936	920	973	921	842	819	714	125	6	3	2
4	4	29	568	868	867	905	909	936	994	954	931	963	875	813	490	15	5	4
4	5	131	754	852	906	958	920	963	923	917	904	951	930	851	716	95	6	3
4	5	229	796	879	924	934	923	962	961	993	993	945	989	867	780	162	6	4
3	7	254	827	879	965	949	960	982	926	918	955	927	984	872	765	204	7	3
4	5	175	808	883	996	951	998	935	976	971	940	922	961	872	804	132	4	4
4	4	57	666	859	968	999	947	977	985	916	928	960	974	841	678	62	4	4
4	3	11	406	839	897	915	930	946	993	914	911	977	900	830	359	10	3	4
3	2	5	60	624	830	890	973	903	921	912	930	881	850	613	54	6	3	3
3	4	4	7	92	602	873	856	882	913	887	885	842	589	82	7	4	3	3
3	4	3	4	5	23	266	697	838	828	837	667	261	21	5	4	4	5	4
3	3	4	4	4	6	9	12	27	49	28	11	9	7	5	3	3	4	3
3	5	3	5	4	4	7	4	4	6	6	3	5	3	3	3	3	4	4

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 $160\mu\text{m}$ circle (Hough Transform)

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

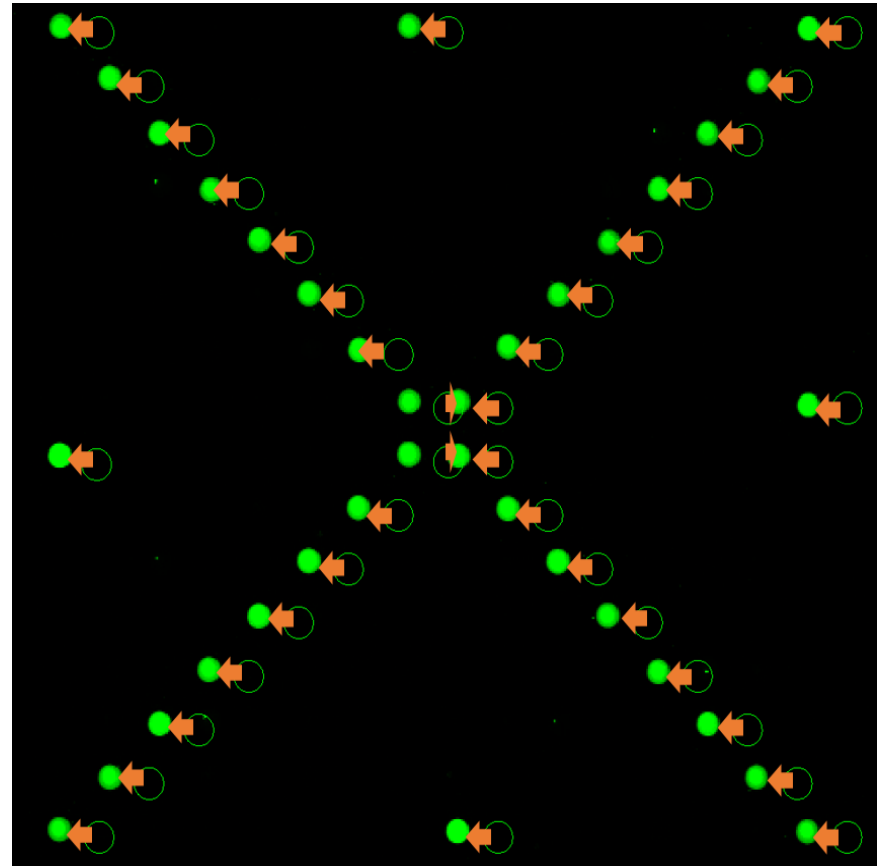


Aligning a GAL file to the Sentinels

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

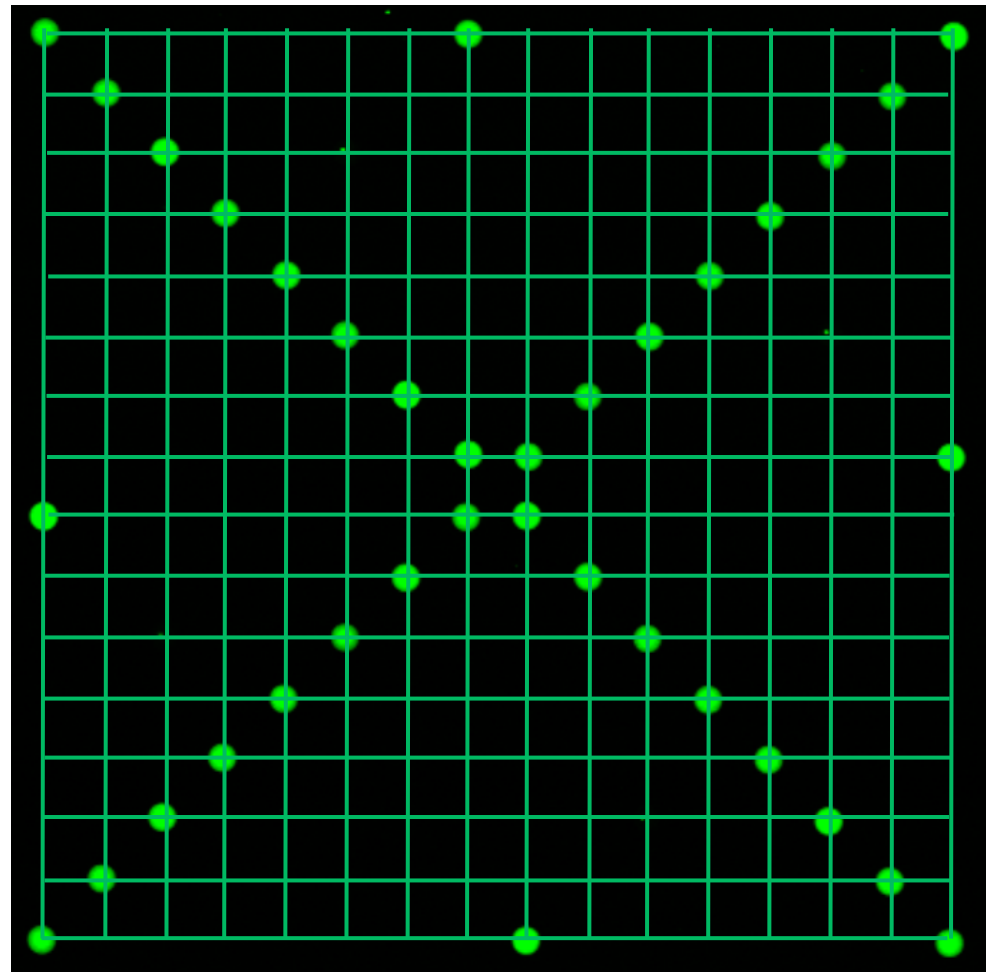


But why so many Sentinels?

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

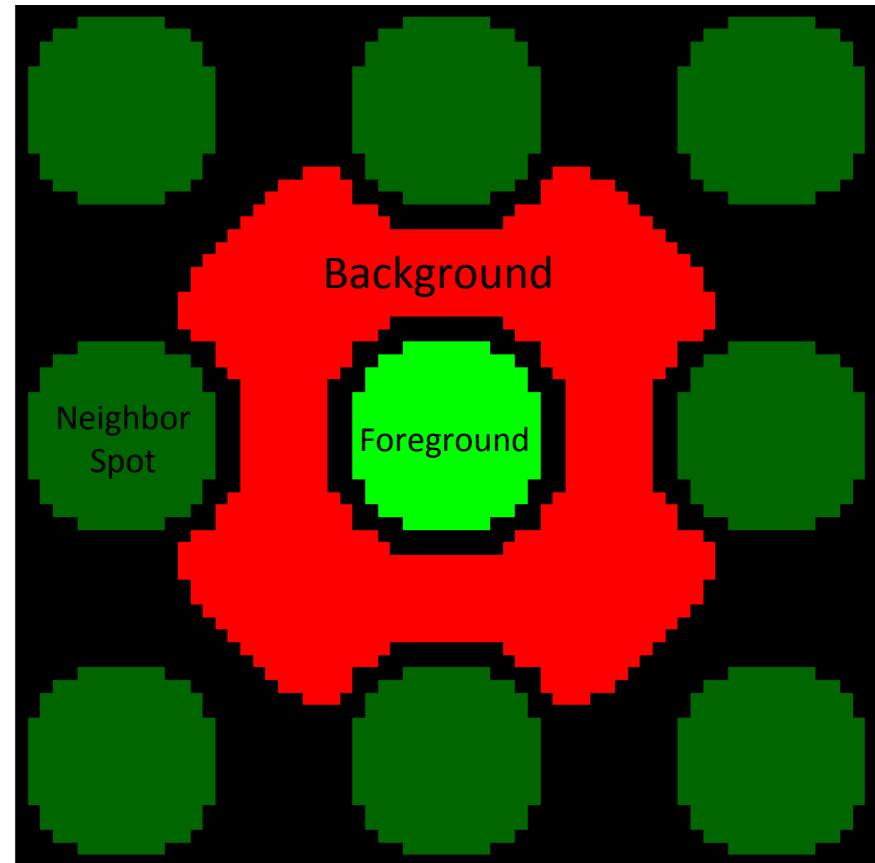
Foreground – Pixels belonging to the printed region, or due to interaction with the compound

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.

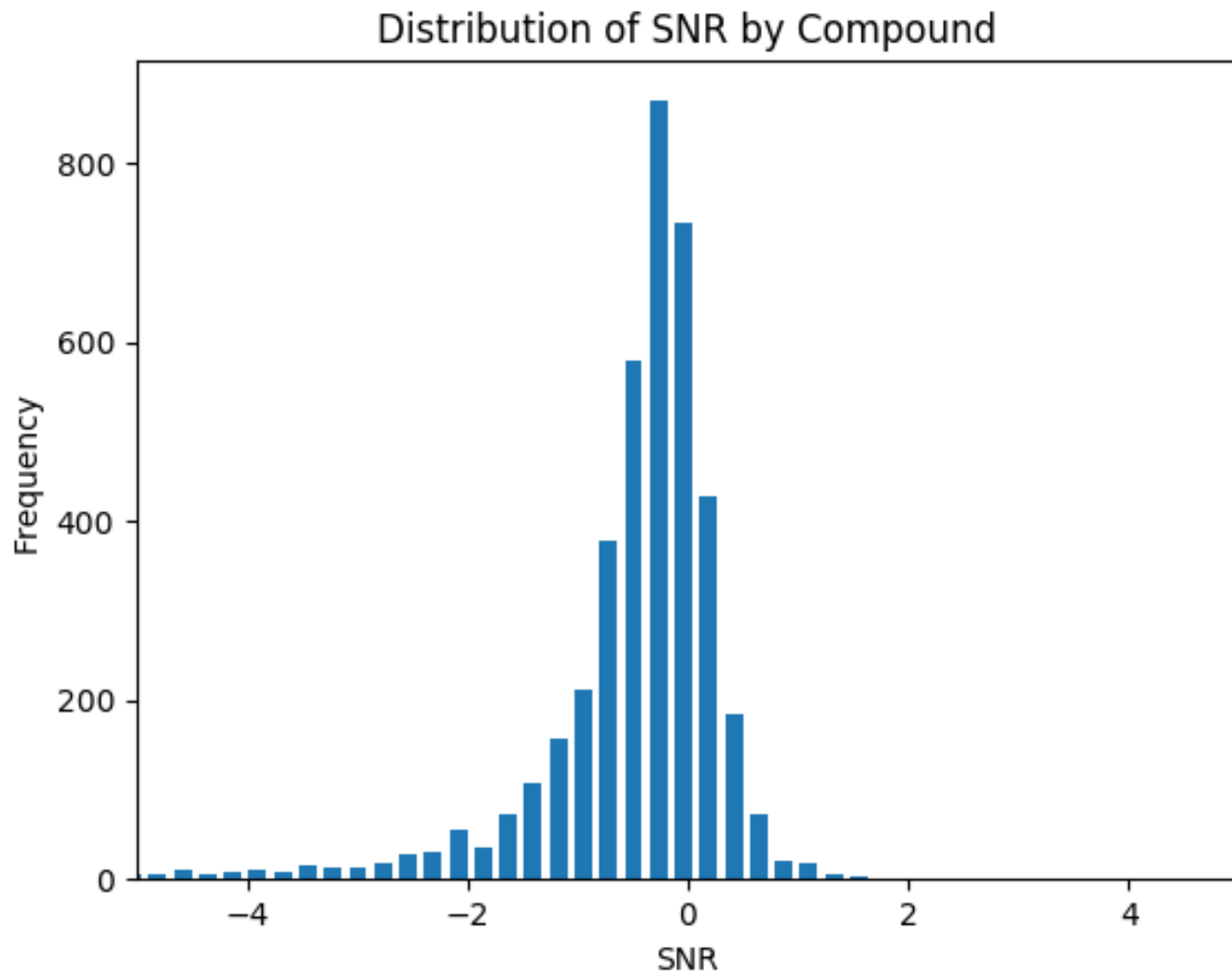


Signal-to-Noise Ratio (SNR)

$$\text{SNR} = \frac{\mu_{\text{foreground}} - \mu_{\text{background}}}{\sigma_{\text{background}}}$$

Signal – Subtract background from the foreground

Noise – Weight by variation in the background



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

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?