20.109 Module 1
Lecture 4
Statistics for High Throughput Science

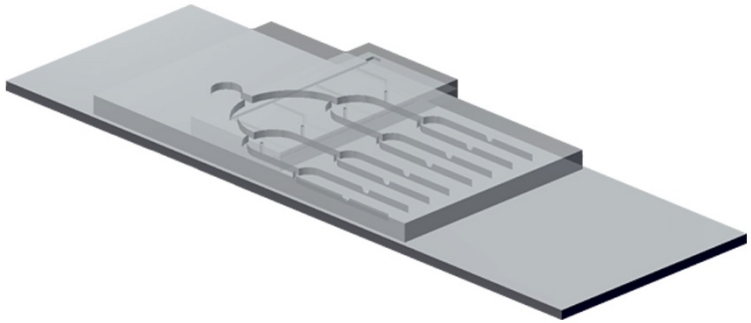
Shelby Doyle
PhD Candidate, Koehler Lab
MIT Biological Engineering

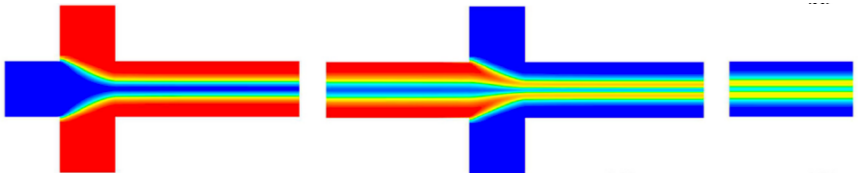
I used to be a sophomore, too!

http://www.lsu.edu/

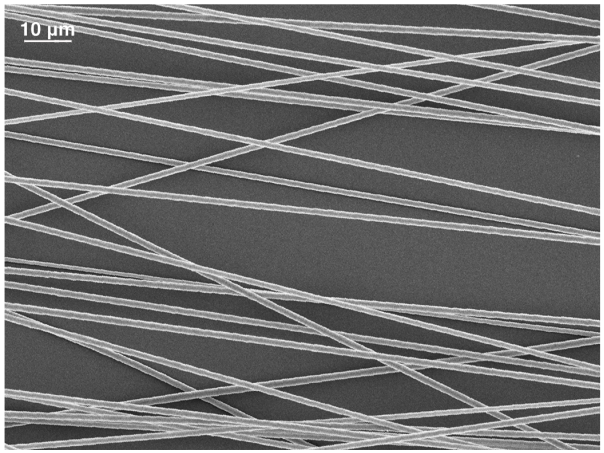
In undergrad, I had all sorts of projects

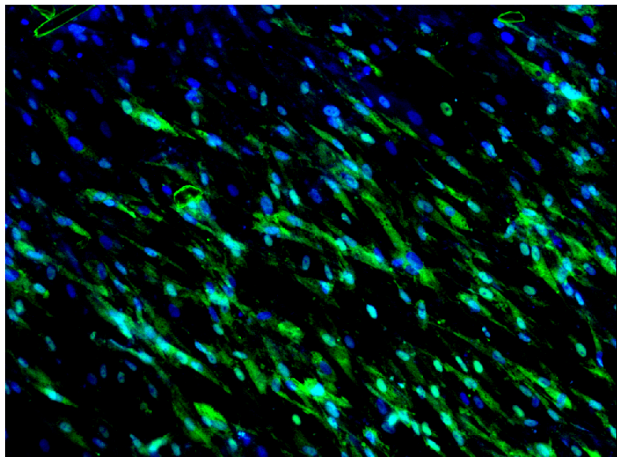
Microfluidics for Drug Testing and Cryopreservation





Materials Science for Tissue Engineering



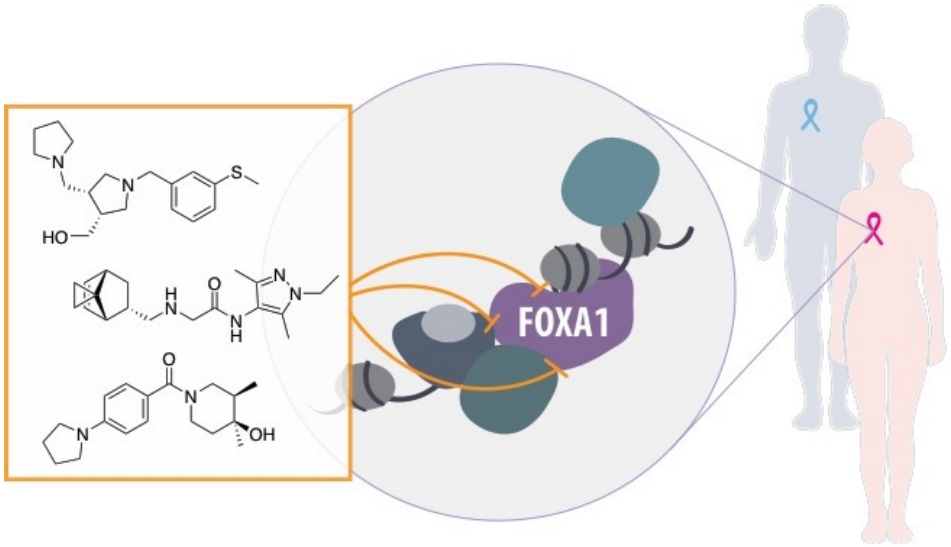


Now, I work with Angela at the Koch Institute

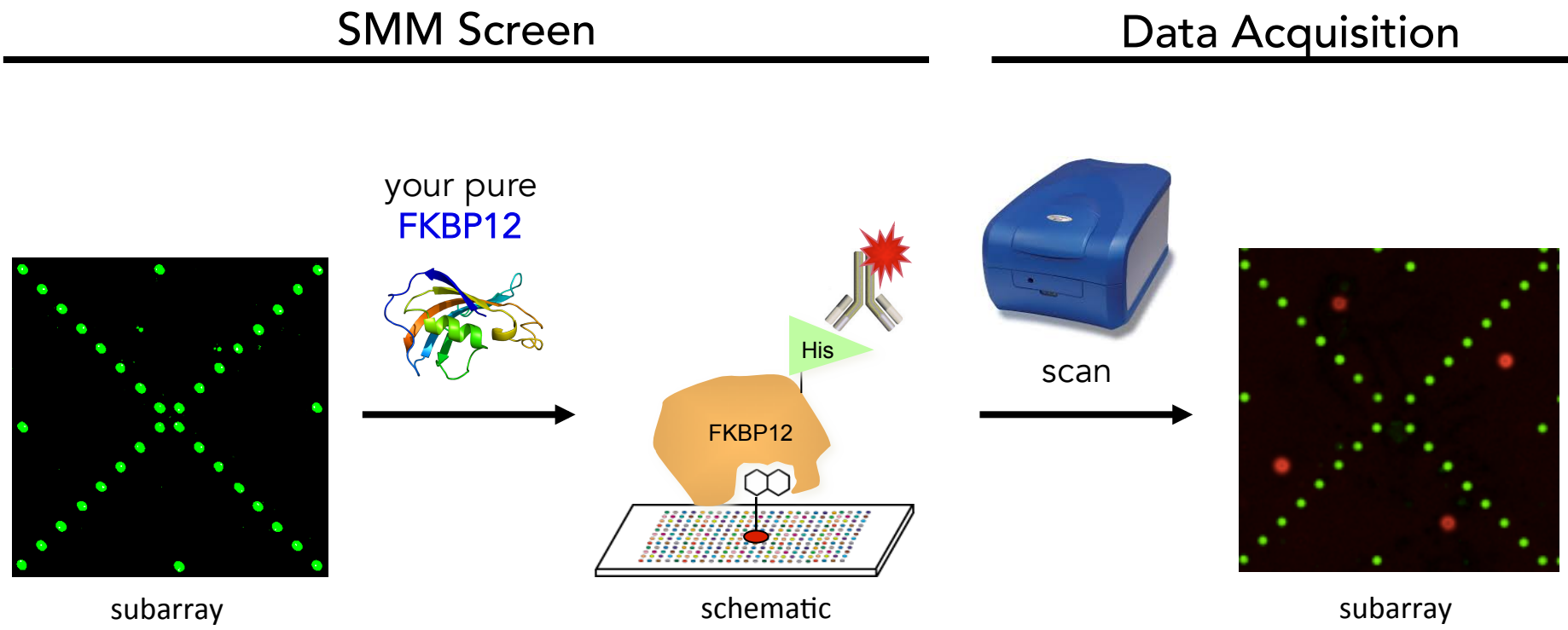
http://capitalprojects.mit.edu/

MIT BE
BIOLOGICAL ENGINEERING

Small Molecule Probe Discovery in Cancer



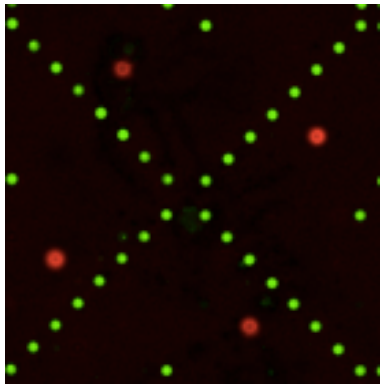
SMM Screening and Data Analysis



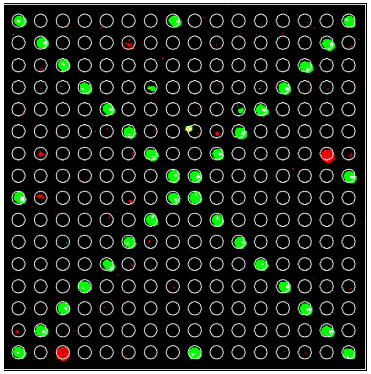
SMM Screening and Data Analysis

Data Analysis

GenePix® Software



subarray

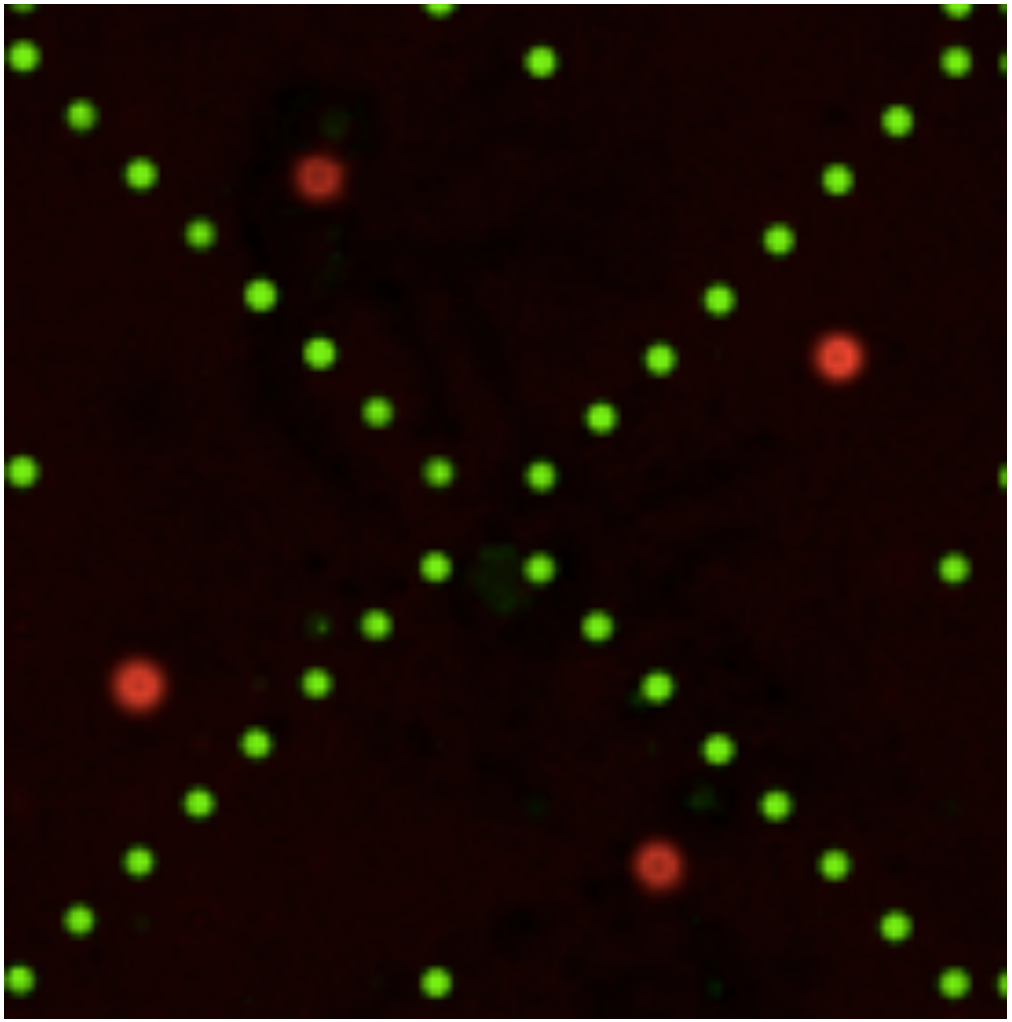


subarray with .gal file overlay

ID	F	B	#	…
Dflt-384-##	#	#	#	
Dflt-384-##	#	#	#	
Dflt-384-##	#	#	#	
Dflt-384-##	#	#	#	
…				

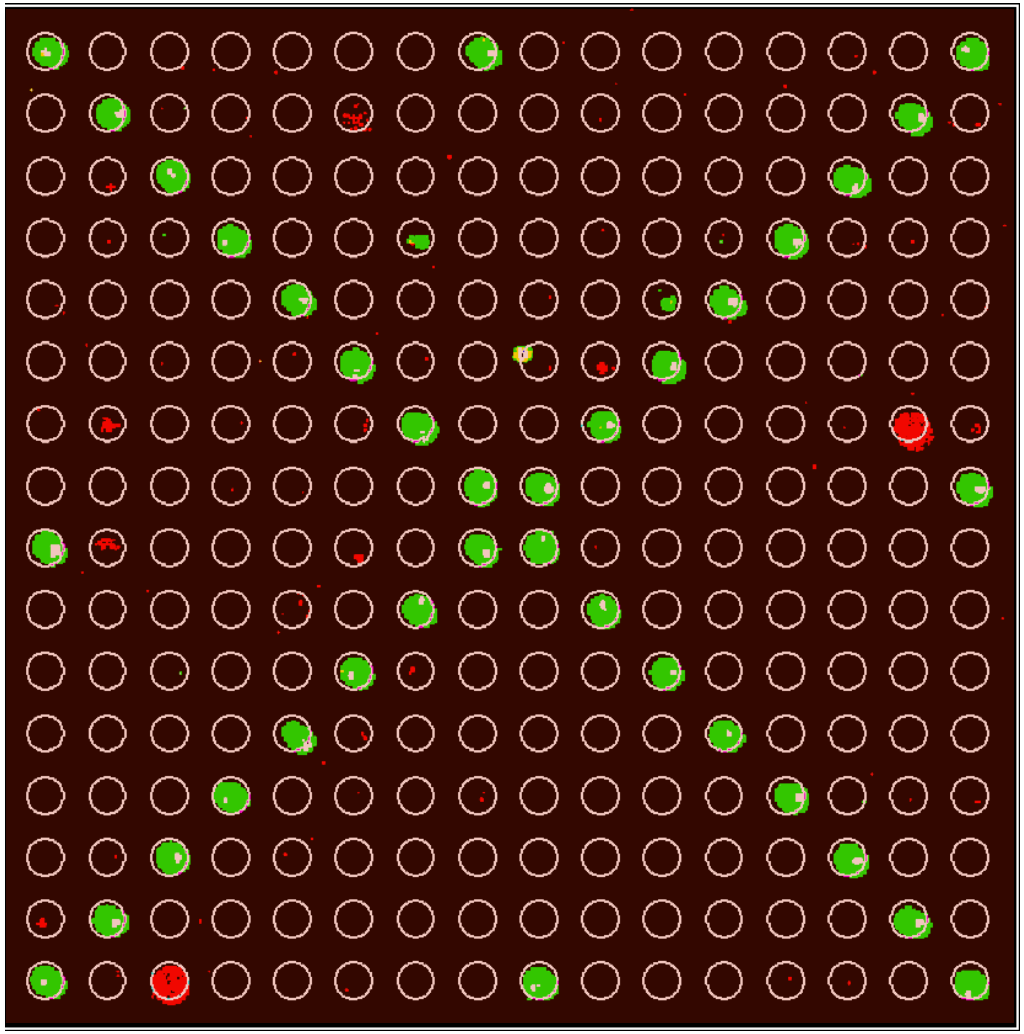
?

How do you know which compounds are binding your protein?



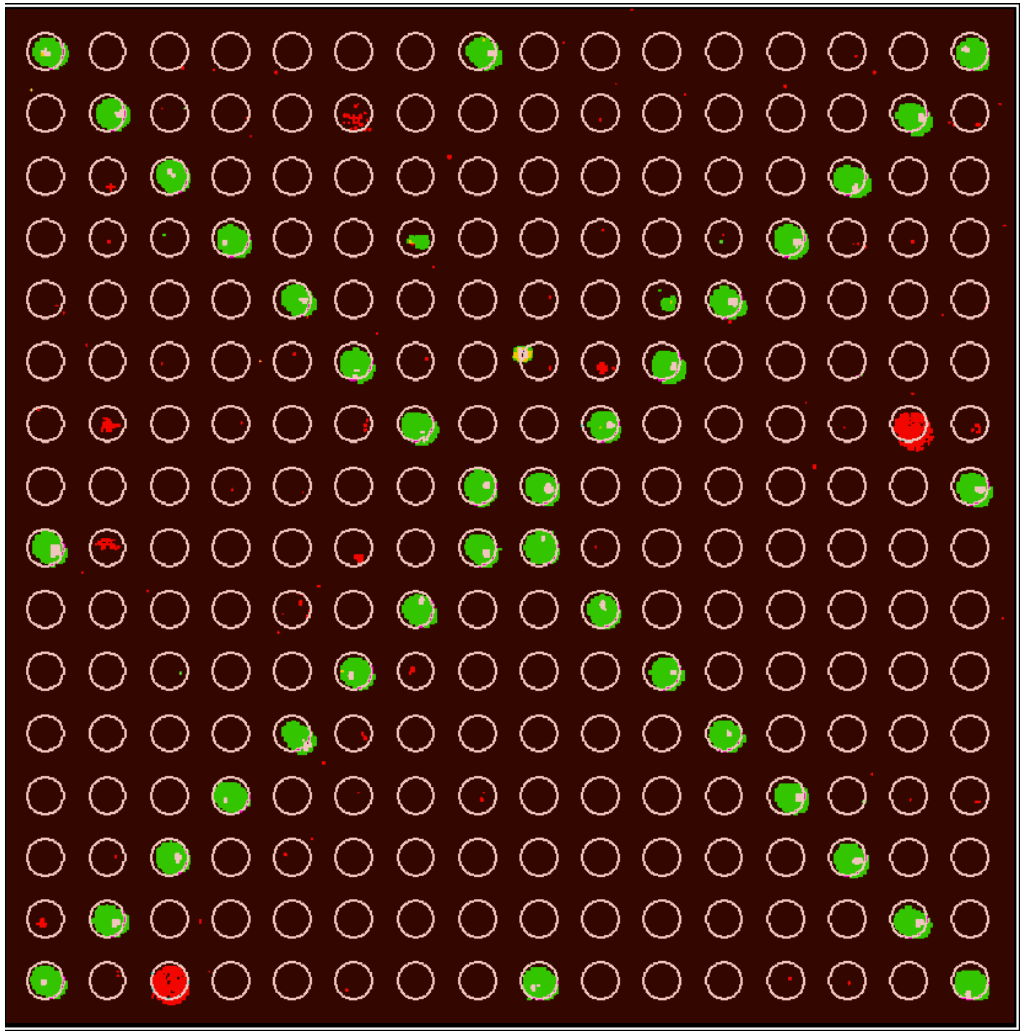
ID	F532	B532	#	…
Dflt-384-##	900	5	#	
Dflt-384-##	5	5	#	
Dflt-384-##	5	5	#	
Dflt-384-##	5	5	#	
…				

How do you know which compounds are binding your protein?



ID	F532	B532	#	…
Dflt-384-##	875	901	#	
Dflt-384-##	893	349	#	
Dflt-384-##	1203	902	#	
Dflt-384-##	403	354	#	
…				

How do you know which compounds are binding your protein?

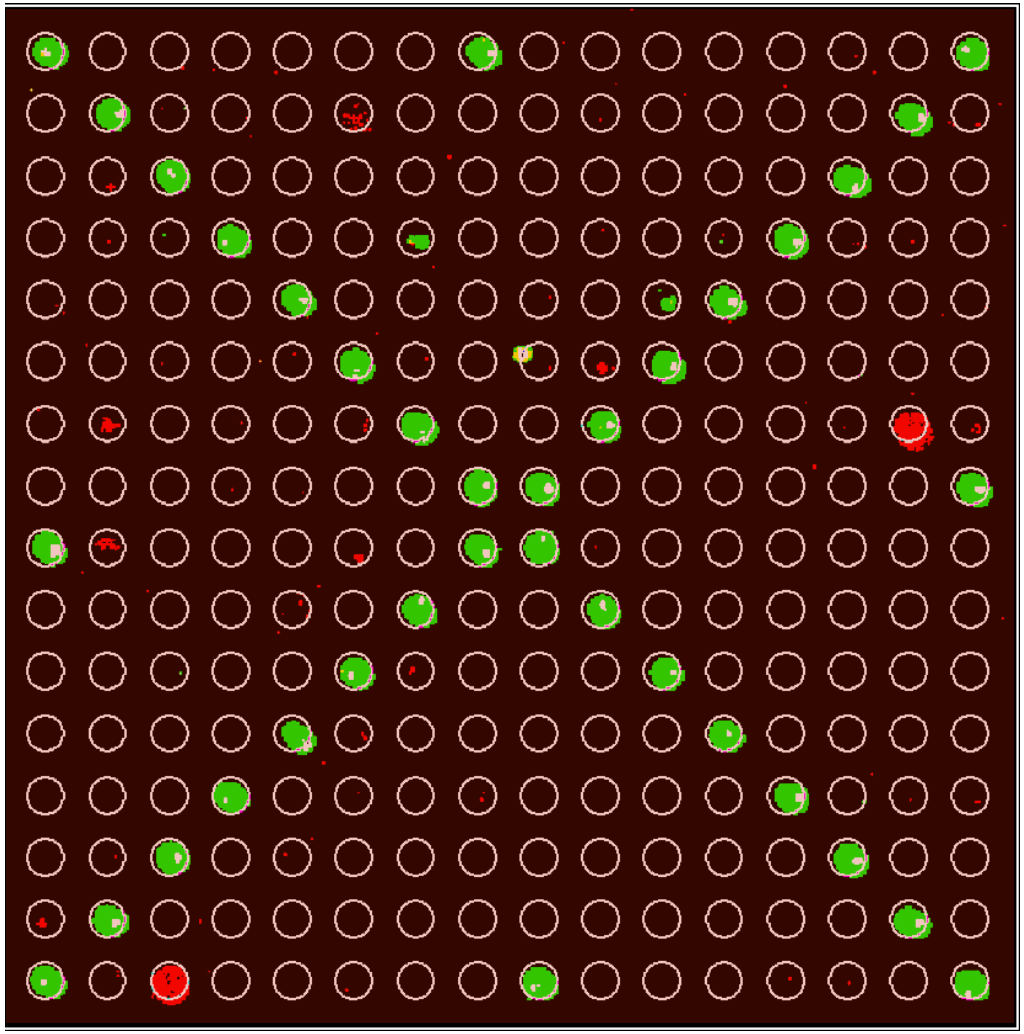


ID	F532	B532	#	…
Dflt-384-##	875	901	#	
Dflt-384-##	893	349	#	
Dflt-384-##	1203	902	#	
Dflt-384-##	403	354	#	
…				

Define a signal to noise ratio (SNR) for each feature

$$SNR\downarrow i = Feature\ Median\ /\ Background\ Median$$

How do you know which compounds are binding your protein?



ID	F532	B532	SNR
Dflt-384-##	875	901	0.97
Dflt-384-##	893	749	1.19
Dflt-384-##	1203	902	1.33
Dflt-384-##	403	354	1.14
…			

But how do you know which compounds are the good ones?!

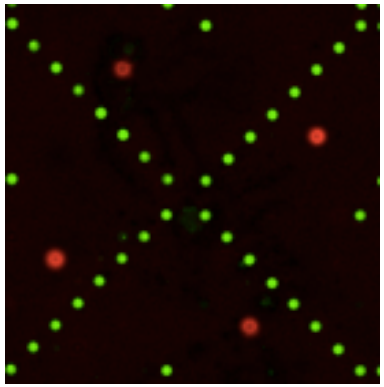
Content and Goals for Today's Lecture

Statistics	Understand the definition of… Know how to calculate a…	Z-score Coefficient of Variation Z-factor
Data Analysis	Be familiar with …	SMM data analysis workflow

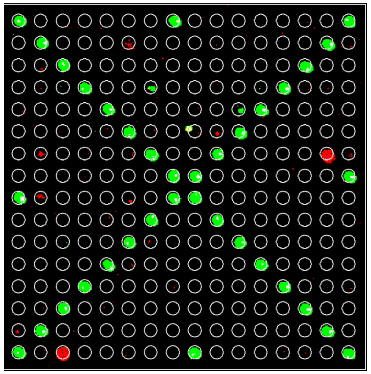
SMM Screening and Data Analysis

Data Analysis

GenePix® Software



subarray



subarray with .gal file overlay

ID	F	B	#	…
Dflt-384-##	..	..	..	
Dflt-3				
Dflt-3				
Dflt-3				
…				

ID	F532	B532	SNR
Dflt-384-##	875	901	0.97
Dflt-384-##	893	749	1.19
Dflt-384-##	1203	902	1.33
Dflt-384-##	403	354	1.14
…			

Z score

Which compounds are above average?

How far above/below average is this compound?

$$z_i = \text{distance from average} / \text{scaling factor}$$

How meaningful is this difference?

Z score

Which compounds are above average?

How far above/below average is this compound?
Calculate the distance from the average

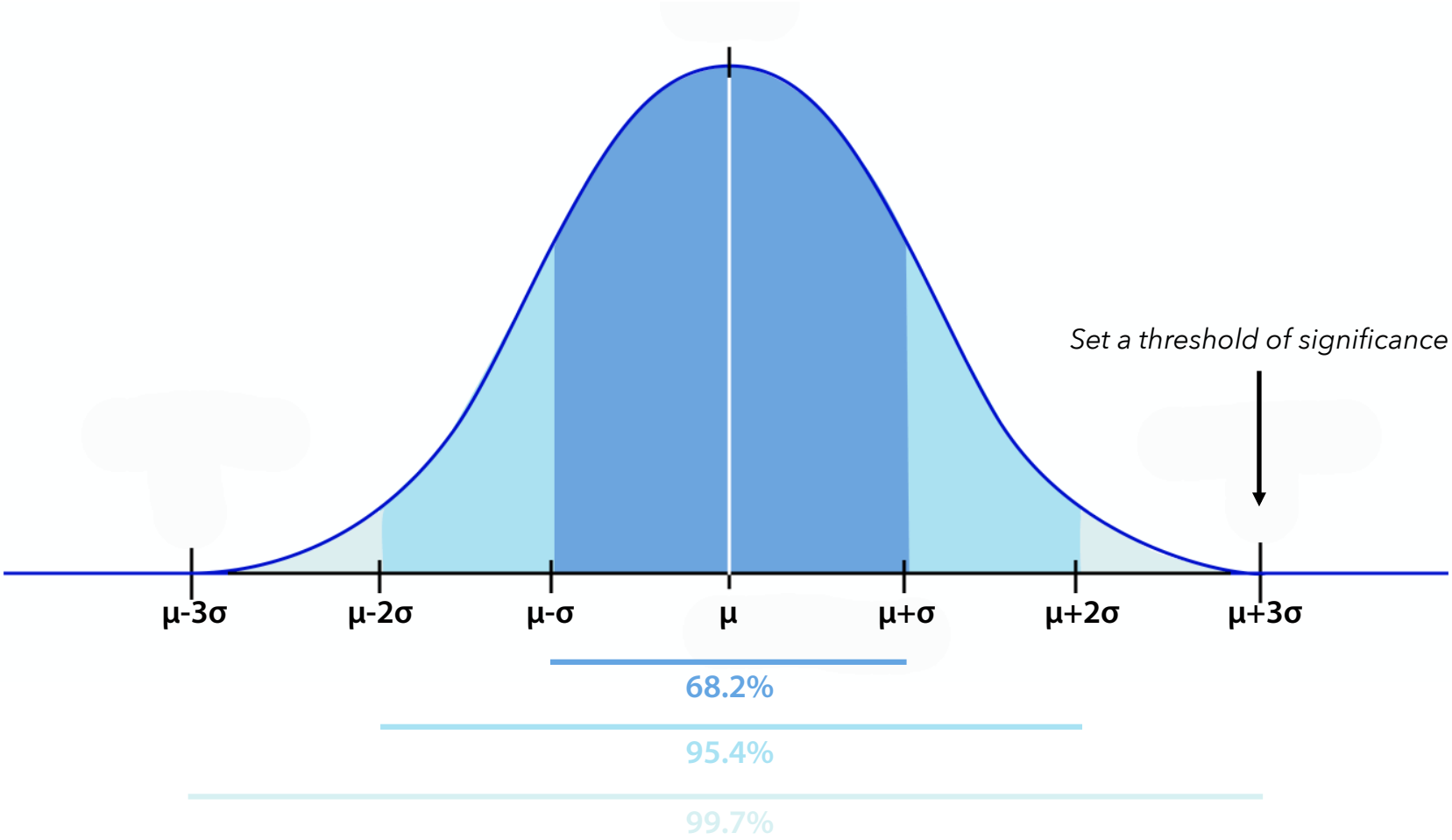
$$z_i = SNR_i - \mu(SNR)/\sigma$$

How meaningful is this difference?
Divide by the standard deviation

$$\sigma = s_{N-1} = \sqrt{1/N-1} \sum_{i=1}^{N} (SNR_i - SNR)^2$$

Hit Calling

What is a good z score?



Note: Not all data is normally distributed

Z score

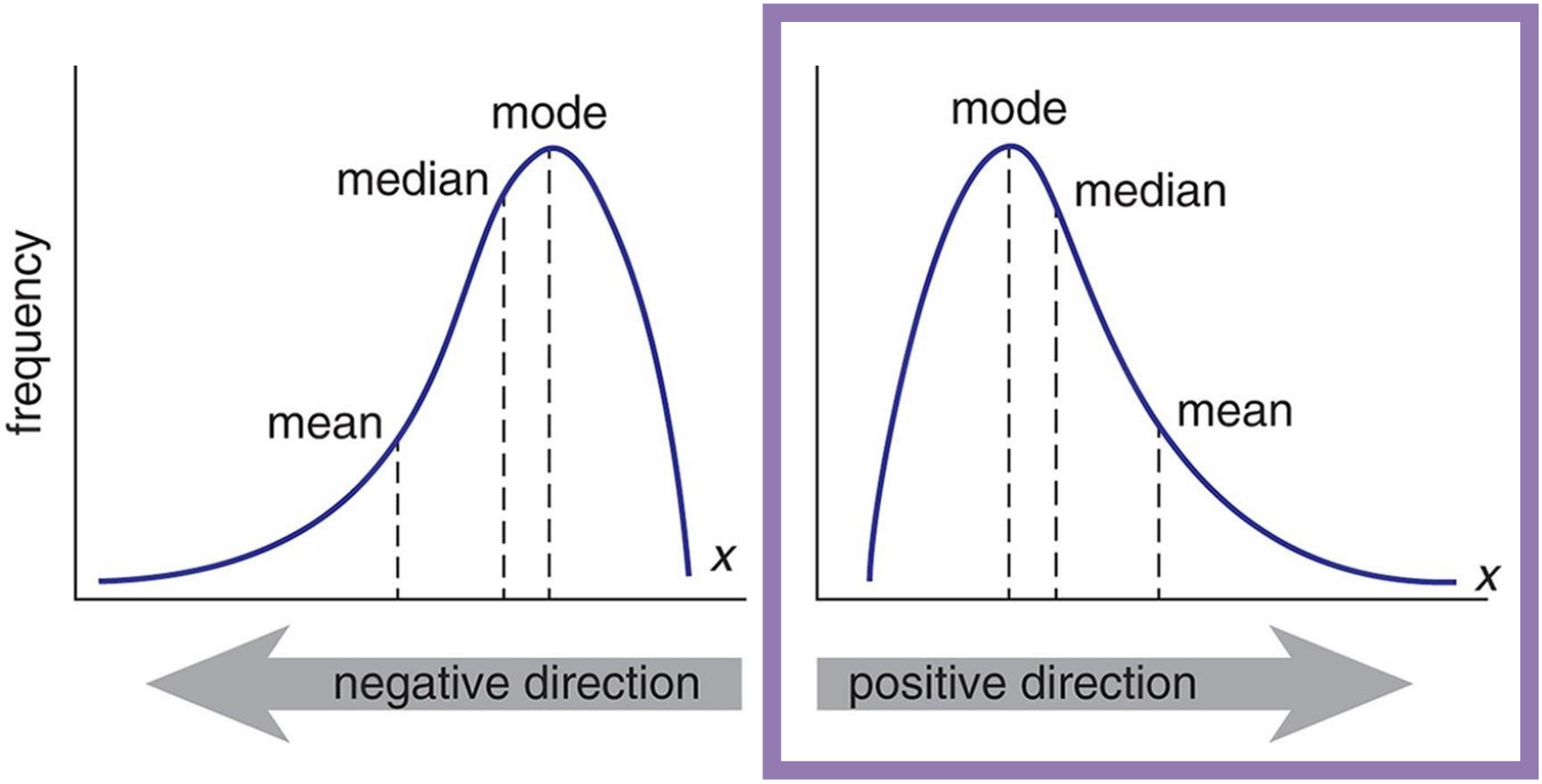
When does it do a bad job?

What if you have something REALLY bright?

$$z\downarrow i = SNR\downarrow i - \mu(SNR)/\sigma$$

Skewedness

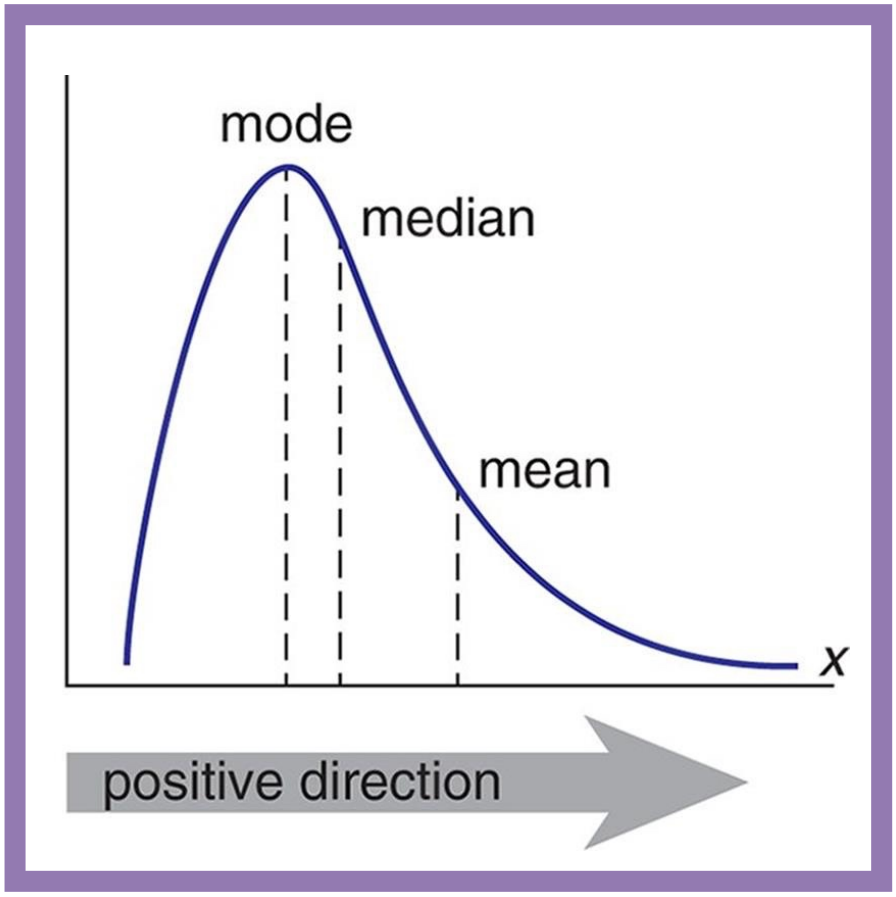
Is the distribution balanced around the mean?



Skewedness

Is the distribution balanced around the mean?

$$z_i = SNR_i - \mu(SNR)/\sigma$$



Robust Z score

Reduce bias due to outliers!

How do we reduce bias in the numerator?
Use the Median instead of the Mean!

$$z_i = SNR_i - Mdn(SNR)/\sigma$$

Robust Z score

Reduce bias due to outliers!

How do we reduce bias in the numerator?
Use the Median instead of the Mean!

$$z_i = SNR_i - Mdn(SNR) / MAD(SNR) * 1.486$$

Redefine the denominator in terms of the median!
Divide by the Median Absolute Deviation (MAD)

$$MAD = Mdn(|x_i - Mdn(x)|)$$

"the median distance from the median"

e.g. When robust z score saved the day!

HA-tagged FOXA1

Expressed and screened in HEK293T
Detected by primary antibody

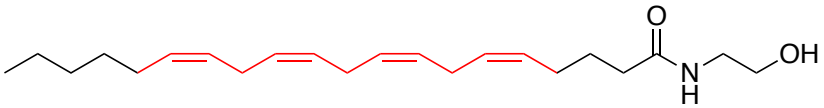
15k Compounds

Diversity Oriented Synthesis Compounds
Known Bioactives

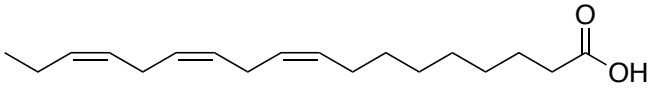
Compound	SNR	Mean	Median	Std Dev	Est. Std Dev	Z Score	Robust Z Score
Anandamide	1.132	1.000	0.998	0.043	0.028	3.052	4.804
	1.123	1.001	0.998	0.041	0.027	2.970	4.669
	1.116	1.001	0.998	0.041	0.027	2.809	4.421
	1.109	1.000	0.998	0.043	0.028	2.517	3.972
Linolenic Acid	1.097	1.000	0.998	0.043	0.028	2.251	3.559
	1.060	1.001	0.998	0.041	0.027	1.453	2.331
	1.032	1.001	0.998	0.041	0.027	0.758	1.258
	1.025	1.000	0.998	0.043	0.028	0.581	0.965

e.g. When robust z score saved the day!

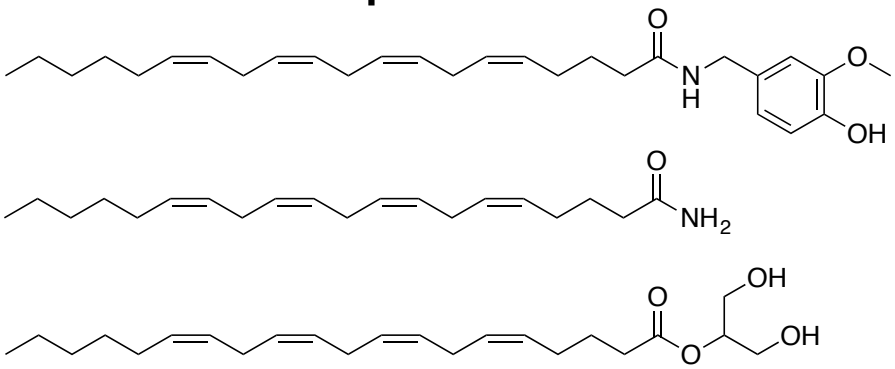
Anandamide



Linolenic Acid



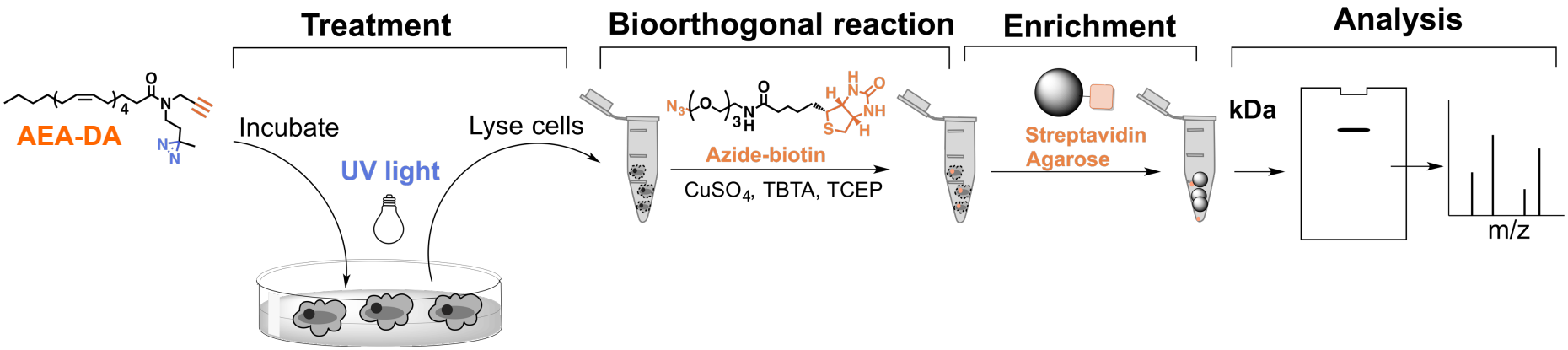
C20:4 Chain Compounds

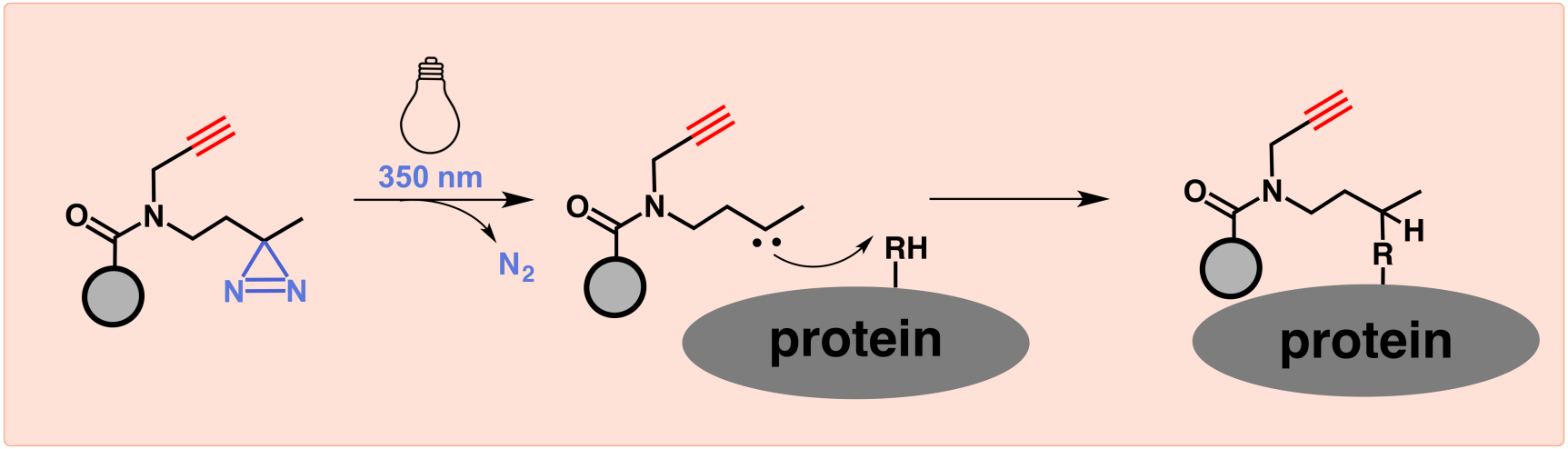


Compound	SNR	Mean	Median	Std Dev	Est. Std Dev	Z Score	Robust Z Score
Anandamide	1.132	1.000	0.998	0.043	0.028	3.052	4.804
	1.123	1.001	0.998	0.041	0.027	2.970	4.669
	1.116	1.001	0.998	0.041	0.027	2.809	4.421
	1.109	1.000	0.998	0.043	0.028	2.517	3.972
Linolenic Acid	1.097	1.000	0.998	0.043	0.028	2.251	3.559
	1.060	1.001	0.998	0.041	0.027	1.453	2.331
	1.032	1.001	0.998	0.041	0.027	0.758	1.258
	1.025	1.000	0.998	0.043	0.028	0.581	0.965

C20:4 Chain Engages FOXA1 in Live Cells

Workflow





Dr. Helen Evans

Content and Goals for Today's Lecture

Statistics	Understand the definition of… Know how to calculate a…	Z-score Coefficient of Variation Z-factor
Data Analysis	Be familiar with …	SMM data analysis workflow

Coefficient of Variation

How much variability is inherent in the assay?

$$cv = \sigma/\mu$$

e.g. SMM Screen of FOXA1

DMSO Features	
μ	1.021
σ	0.053
cv	0.052

cv ≤ 0.15

Z factor

How significant is the output of the assay?

The deviation of the positive (pos) and negative (neg) controls

$$Z\prime = 1-3(\sigma_{pos}+\sigma_{neg})/|\mu_{pos}-\mu_{neg}|$$

The difference between the positive (pos) and negative (neg) controls

Z' ≥ 0.5

These statistics are useful for any High Throughput Science approach

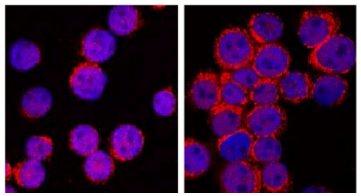
?

e.g. SMM

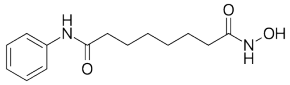
Approaches to probe discovery

'forward' chemical genetics

screen for phenotype of interest

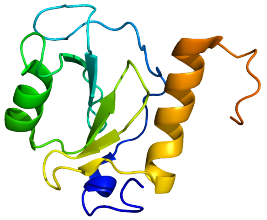


- small molecule + small molecule



assay positive

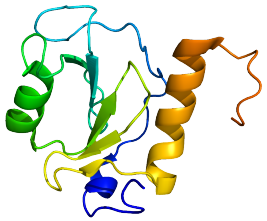
identify protein target

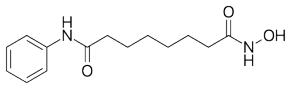


'reverse' chemical genetics

our approach for this class

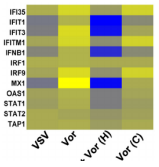
directly bind target of interest

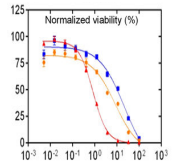


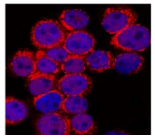


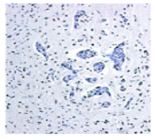
assay positive

broad survey of phenotypic outcomes







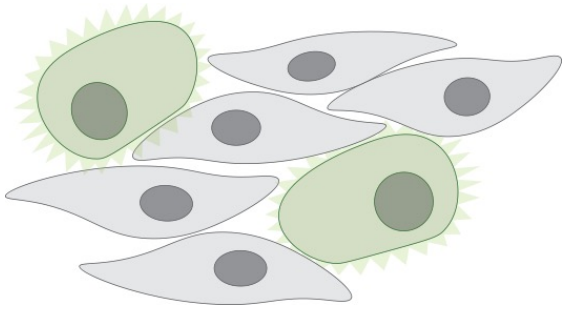


These statistics are useful for any High Throughput Science approach

e.g. Screen for overcoming stromal resistance in Multiple Myeloma (MM)

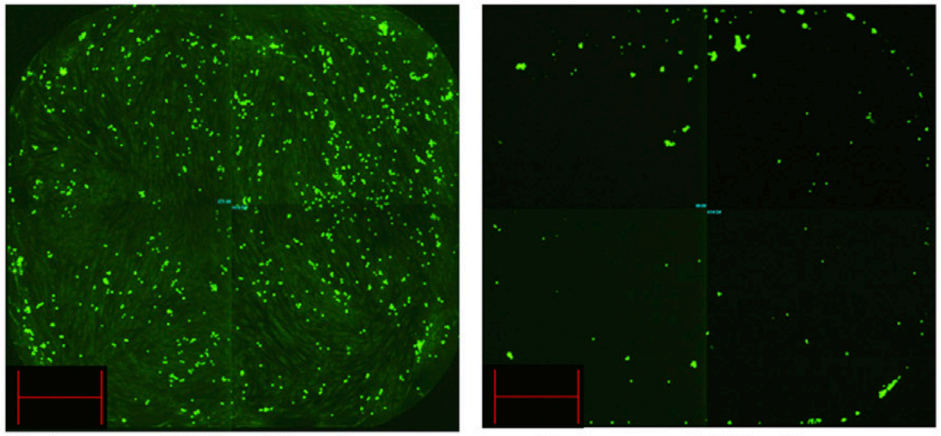
Live Cell Screen

Bone marrow stromal cells



GFP-expressing multiple myeloma cells

Data Acquisition



Primary MM with stroma **Primary MM alone**

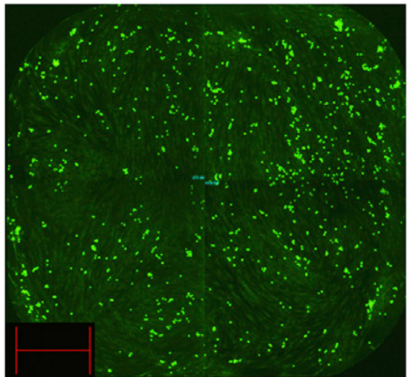
Chattopadhyay, S. et al. (2015). Cell Reports, 10(5), 755–770.

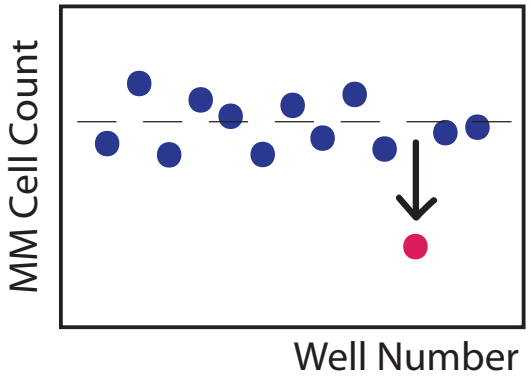
These statistics are useful for any High Throughput Science approach

e.g. Screen for overcoming stromal resistance in Multiple Myeloma (MM)

Data Analysis

CellProfiler
cell image analysis software





24,320 cpds

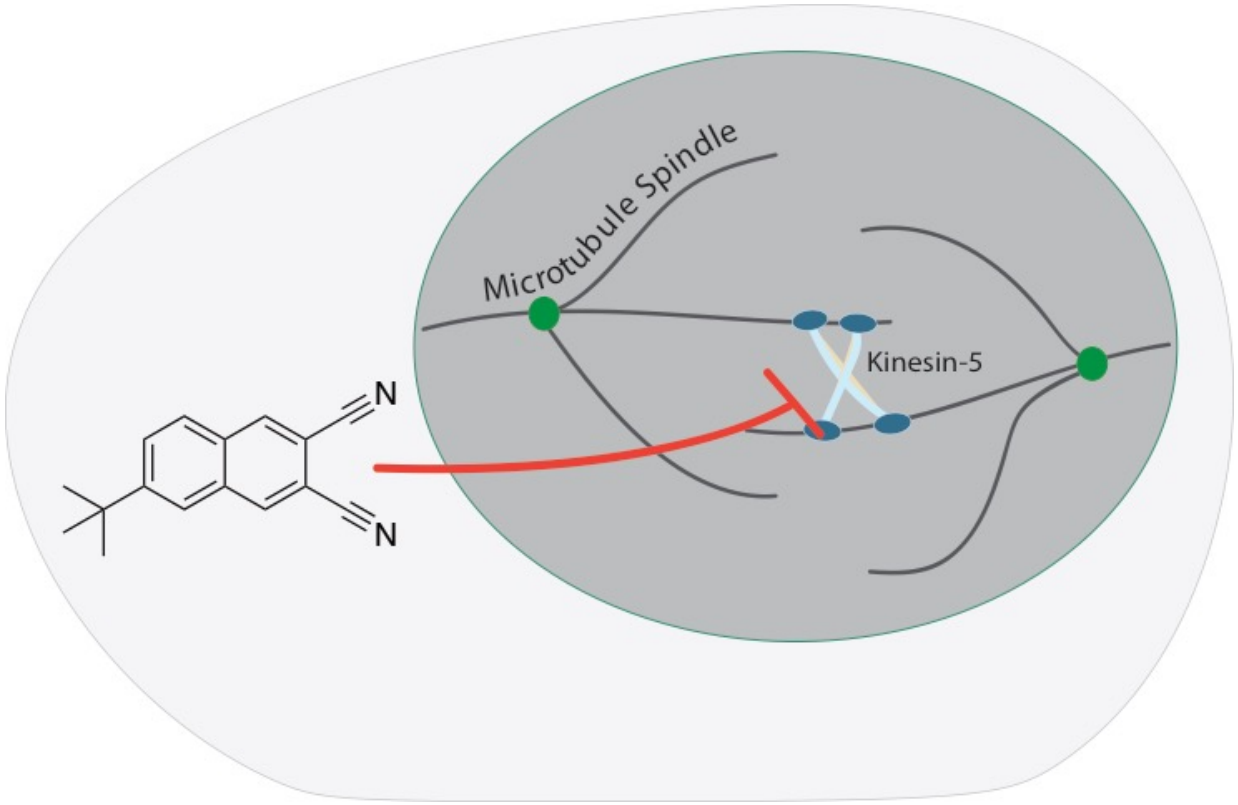
790 hits

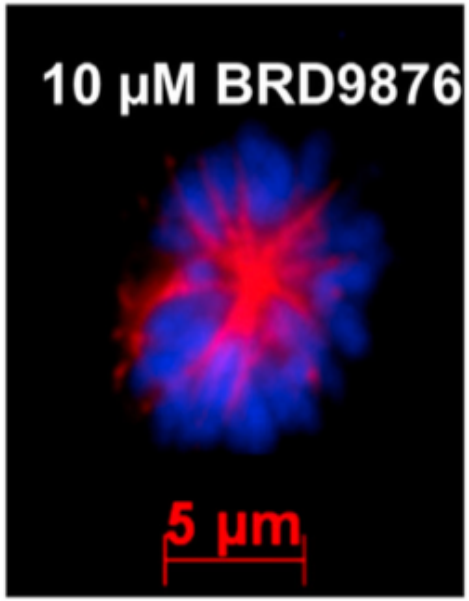
Z′ ≥ 0.5
Z score ≤ - 2

Chattopadhyay, S. et al. (2015). Cell Reports, 10(5), 755–770.

These statistics are useful for any High Throughput Science approach

e.g. Screen for overcoming stromal resistance in Multiple Myeloma (MM)





Chattopadhyay, S. et al. (2015). Cell Reports, 10(5), 755–770.

Our path to probe discovery - lectures

2/14/17	Lecture 1	Intro to chemical biology: small molecules, probes, and screens
2/16/17	Lecture 2	For the love of proteins: FKBP12 and immunophilins
2/21/17	No Lecture	
2/23/17	Lecture 3	Small-molecule microarrays
2/28/17	Lecture 4	Analyzing SMM data sets (Shelby Doyle)
3/2/17	**Lecture 5**	**Chemical probe stories**
3/7/17	Lecture 6	Wrap up discussion: suggestions for how to report your findings