7-17-2019

The statistics of directed evolution: From library generation to high throughput screens

Keith Tyo

THE STATISTICS OF DIRECTED EVOLUTION

FROM LIBRARY GENERATION TO HIGH THROUGHPUT SCREENS

Keith EJ Tyo

Workshop materials
https://bit.ly/2SnRoEf

Workshop goals

- Understand the role of noise in high throughput screens
- Devise screening strategies that are robust to noise
- Understand the effects of different diversification strategy

Northwestern CENTER FOR SYNTHETIC BIOLOGY™

Northwestern McCORMICK SCHOOL OF ENGINEERING

Disclaimer

- Ask questions
- Offer alternate approaches

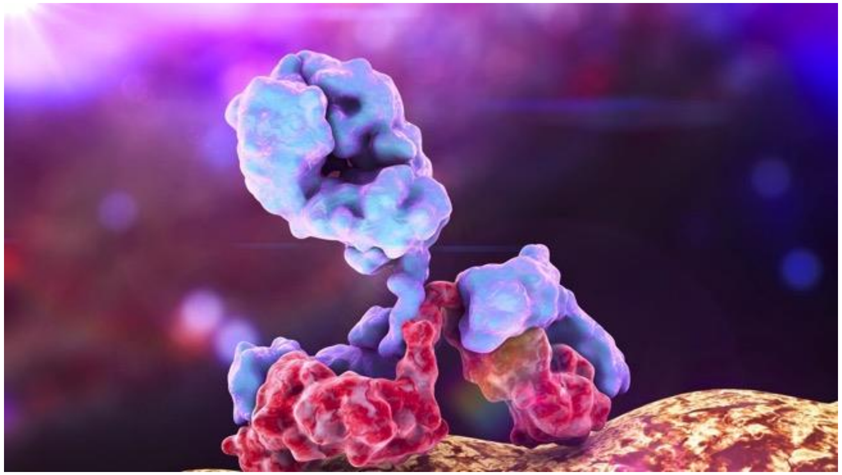
Workshop materials

https://bit.ly/2SnRoEf

Directed evolution as a tool to improve traits

Biological parts

- Enzymes
- Antibodies
- Metabolic networks
- Cells



Traits

- Activity
- Specificity
- Stability
- Affinity
- Metabolic regulation
- Stress tolerance

Directed evolution impact

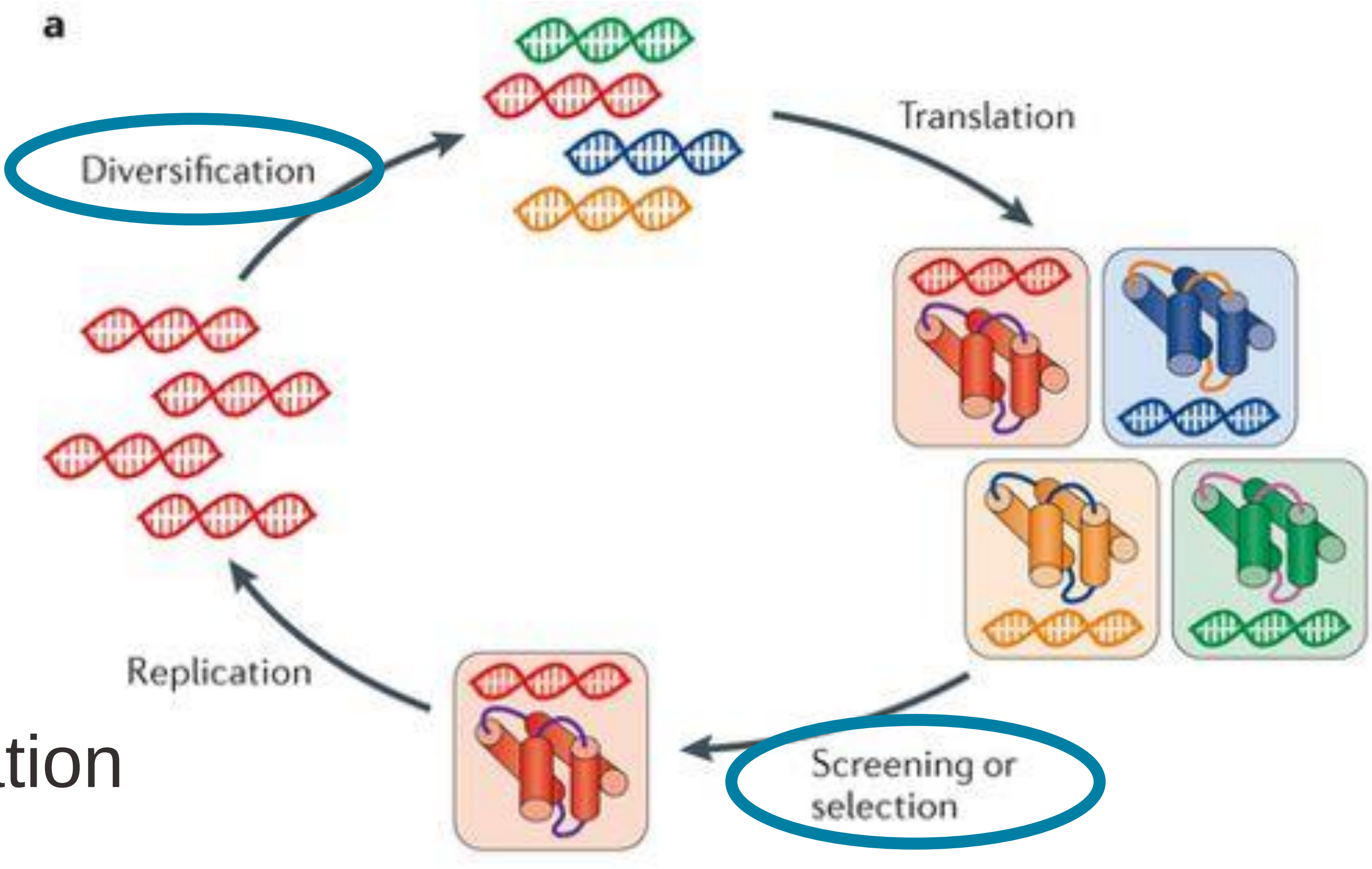
Industries

- Chemical
- Agriculture
- Therapeutics
- Diagnostics

© Nobel Media AB.
Photo: A. Mahmoud

2018 Nobel Prize
in Chemistry

Directed evolution process



Topics

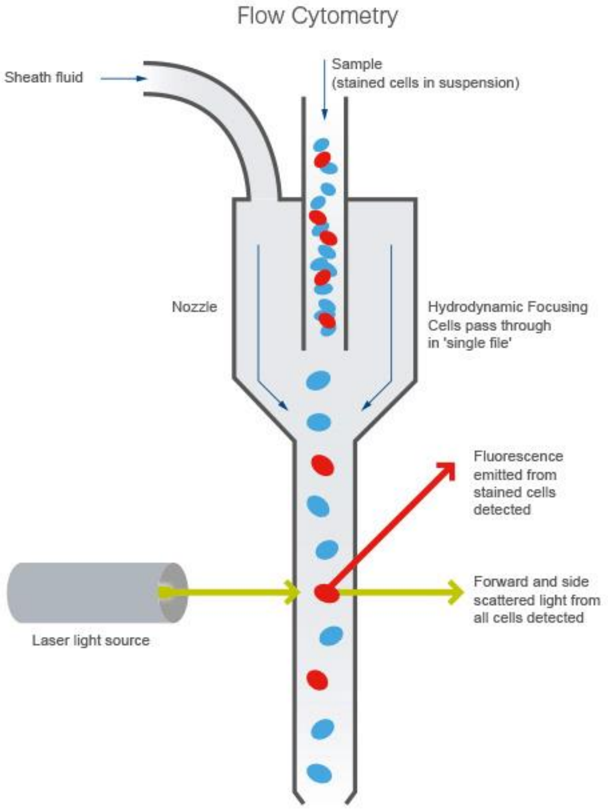
1. Screening
2. Diversification

Packer, M. S., & Liu, D. R. (2015). *Nature Reviews. Genetics*.

SCREENING

Screening in directed evolution

- Flow cytometry
- Microdroplet screening
- 96 well plates
- Affinity separations
  - Phage panning



<u>Proving a negative result</u>

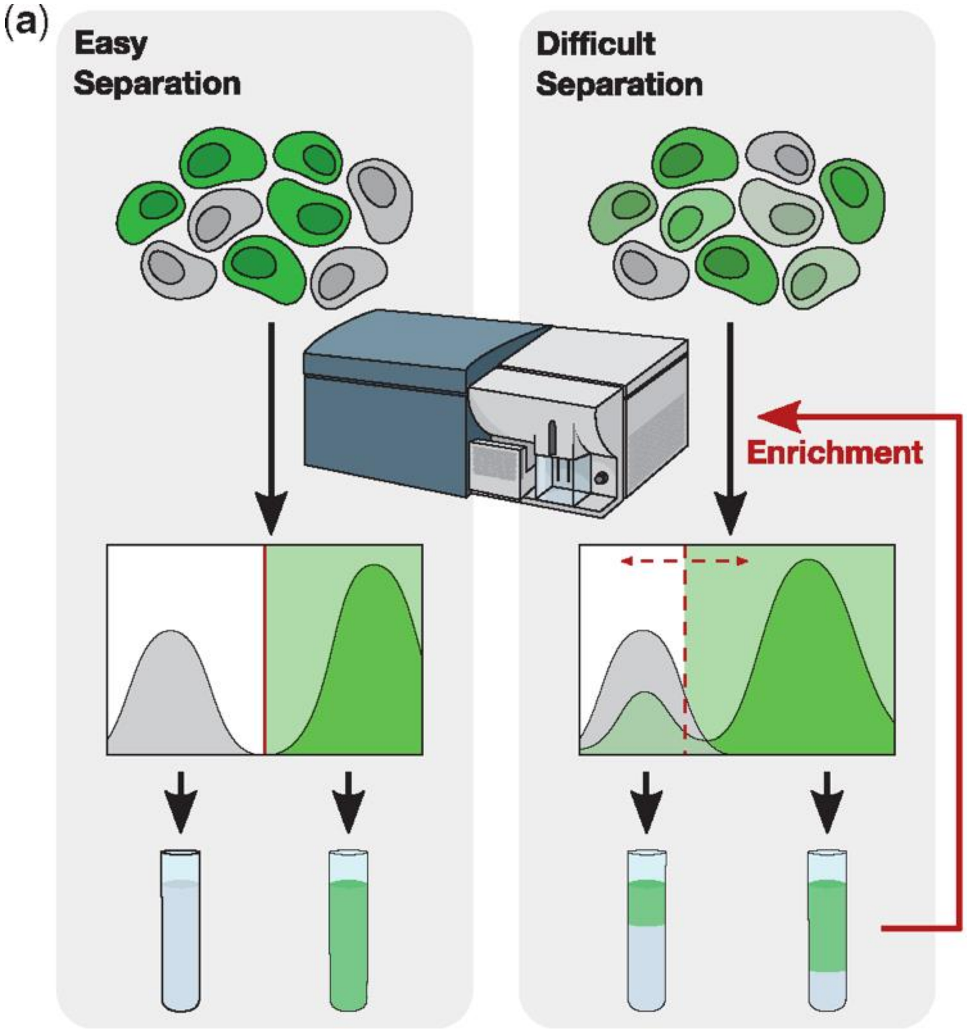
If you do not find an improved mutant …

a. there is no improved mutant in the library

or

b. the screen could not find the improved mutant

Screening

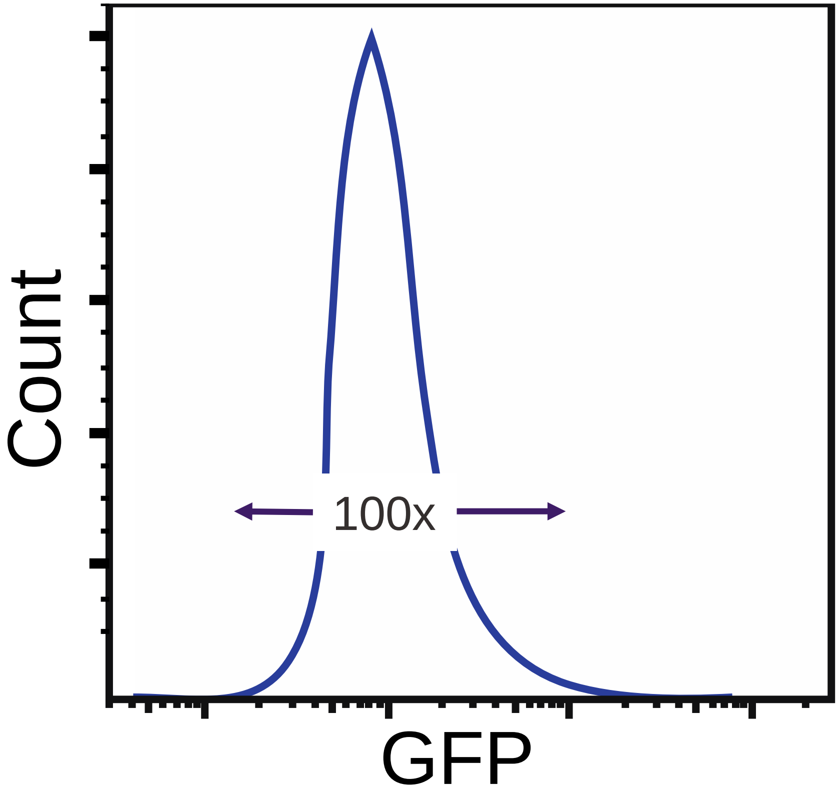


Yu, J. S., Pertusi, D. A., Adeniran, A. V., Tyo, K. E. J., (2017). CellSort:. *Bioinformatics*, *33*(6), 909–916.

Screening

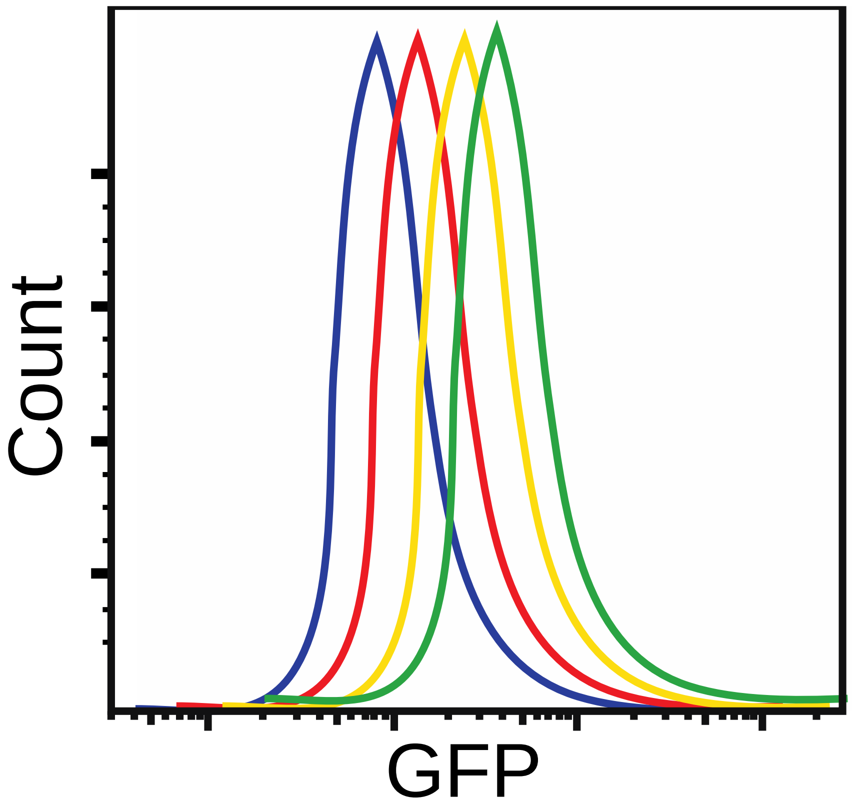
- Single cell measurements are noisy
  - Often 1-2 orders of magnitude
- How big of a shift are you expecting?
  - 50%
- FACS, microfluidic droplets

Clonal populations present a range of values

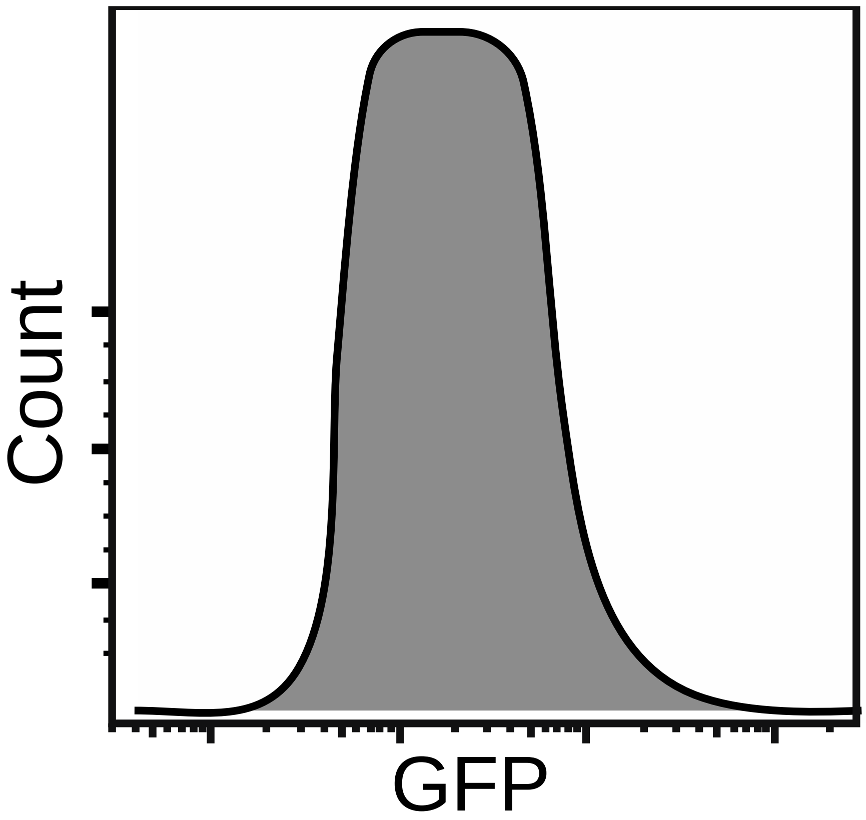


Libraries often create diversity that is smaller than clonal spread

Library of four mutants



Library of four mutants



Depending on where you set your gate, you could still get cells from the poorest mutant

Key questions

- Where should I set my gate?

- How many rounds of enrichment do I need?

Definitions

- WT – wild-type properties

  We will assume most of the library is similar to WT
- Hits – mutants with improved properties

  We will assume there is only one improved mutant in our library.

  i.e. worst-case scenario

Do you have a positive control?

  - i.e. what you expect the improved mutant to look like?

- YES
  - Use flow cytometry plots to find true positive, false negative, etc.

- NO
  - Make assumption about what an improved mutant looks like

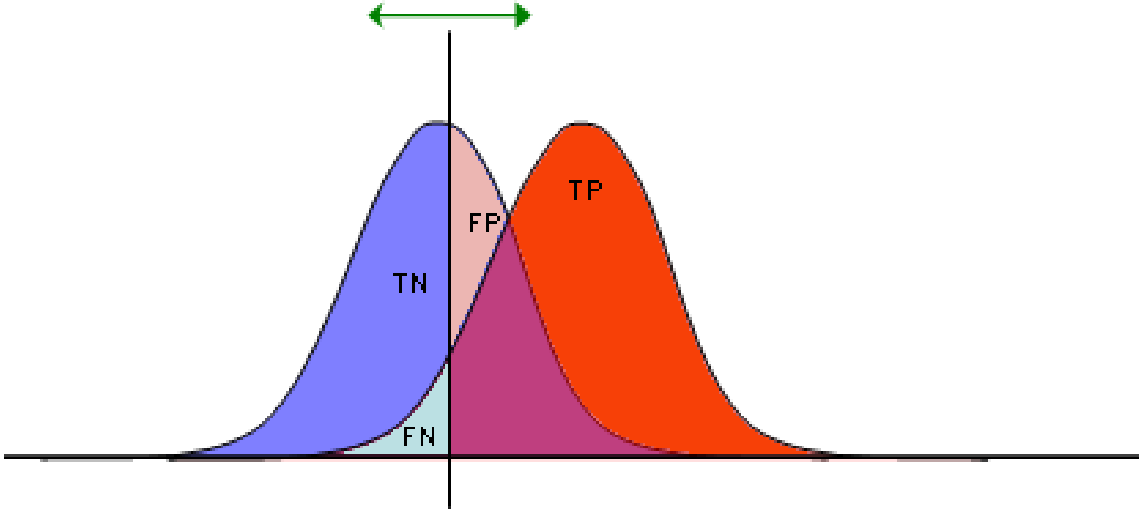
I do have a positive control.

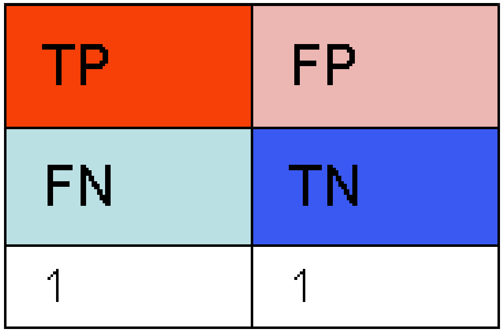
- I have a known binder
  - Neg. control – a non-binding molecule
  - Pos. control - a known binding molecule

- I have a native substrate for an enzyme
  - Neg. control – the new substrate (that does not currently work)
  - Pos. control – the native substrate

- Other

Calculating pos/neg from flow cytometry data





TP	FP
FN	TN
1	1

True positive frequency and false positive frequency can be estimated directly from flow cytometry data.

1. Set threshold
2. Calculate fraction of positive and negative control above the threshold.

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Enrichment

For your screen

- How has the fraction of hits increased after one round of sorting?

$$Enrichment = \frac{Fraction\ of\ positive\ cells\ that\ are\ sorted\ (true\ pos)}{Fraction\ of\ naïve\ library\ that\ is\ sorted\ (true + false\ pos)}$$

How do we know how many improved mutants are in the naïve library?

We don’t.
But, we can assume the worst case scenario.
There is only one improved mutant in the entire library.

Frac of cell= (Library size)$^{-1}$

Multiple rounds of sorting

$$Final\ hit\ freq = (Enrichment)^n \times\ Initial\ hit\ freq$$

where n = number of rounds of sorting

Don’t rely completely on modeling!

If you have strains that are positive and negative controls…

Create mock libraries

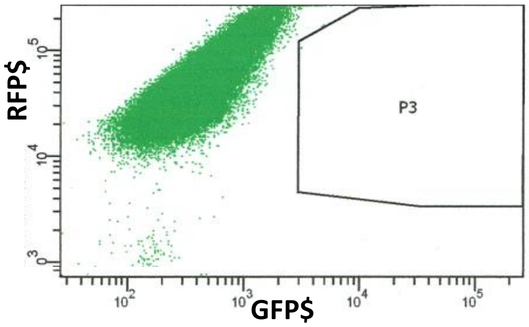
1:100

1:100,000

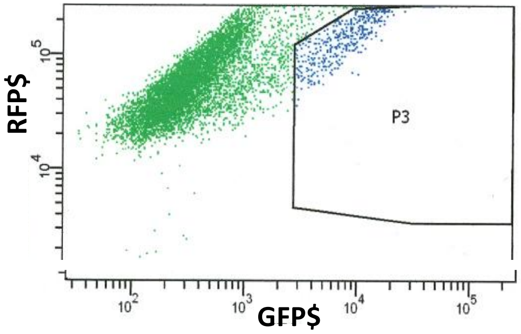
1:1,000,000

Convince yourself that you can recover a rare positive using your procedure

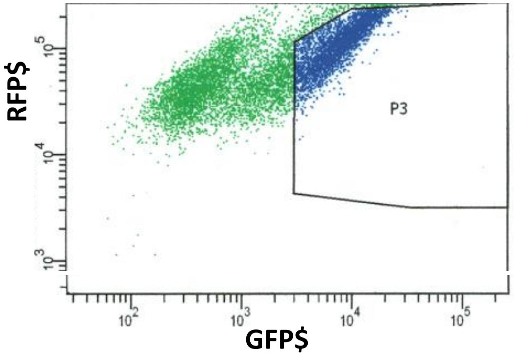
Round 1



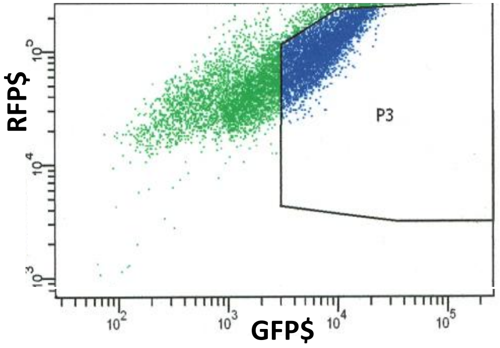
Round 2



Round 3



Round 4



What to do if you don’t have a positive control.

1. Think much harder, and try to find a positive control.

   or

2. Make an assumption about what an improvement might look like.

   Typically coefficient of variance does not change significantly.

   What increase in the mean is reasonable? 10%, 2x, 10x?

Flow cytometry data is typically a log-normal distribution.

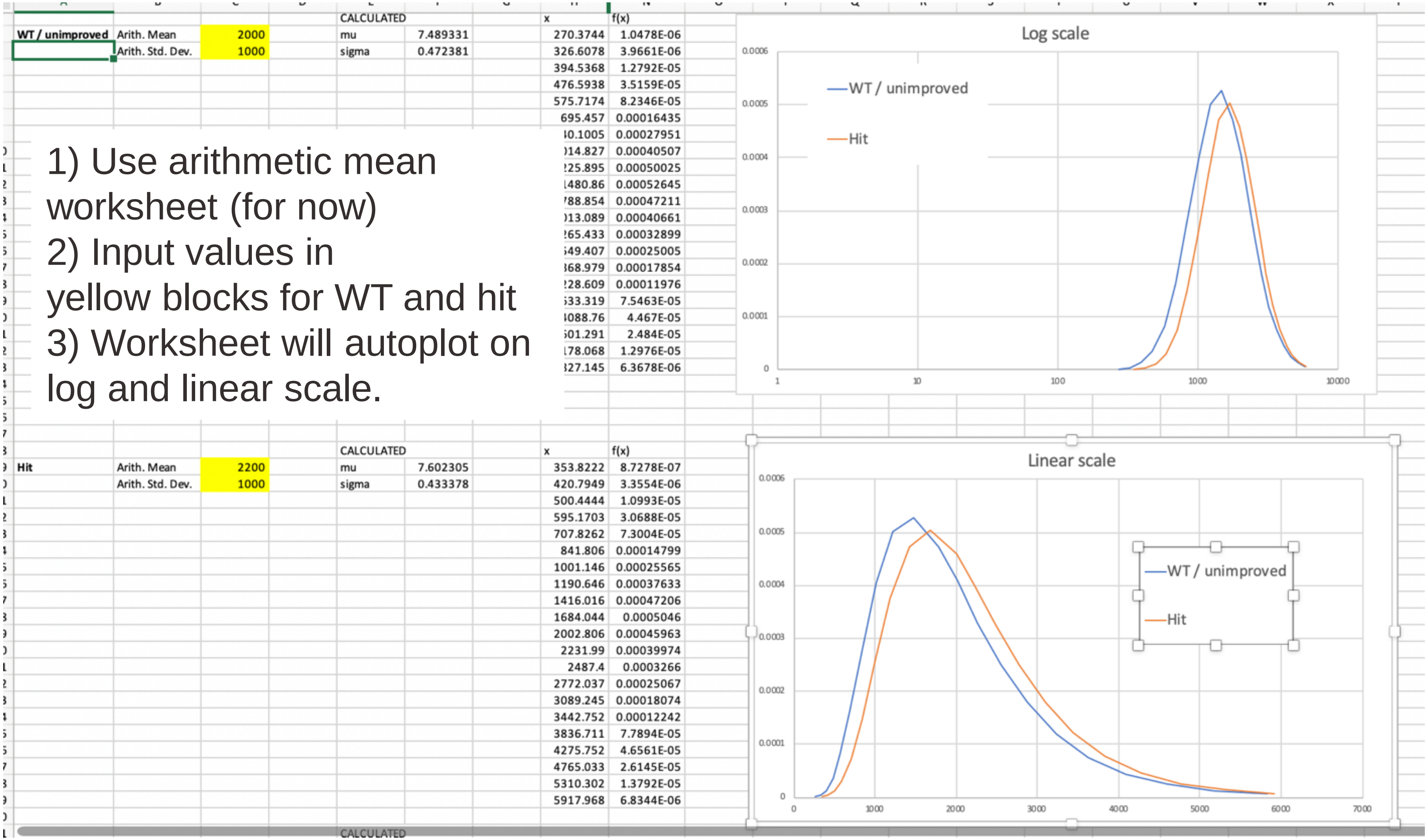
Worksheet was designed for log-normal and normal-normal distributions.

Workshop materials

https://bit.ly/2SnRoEf

Walkthrough the worksheet

https://bit.ly/2SnRoEf



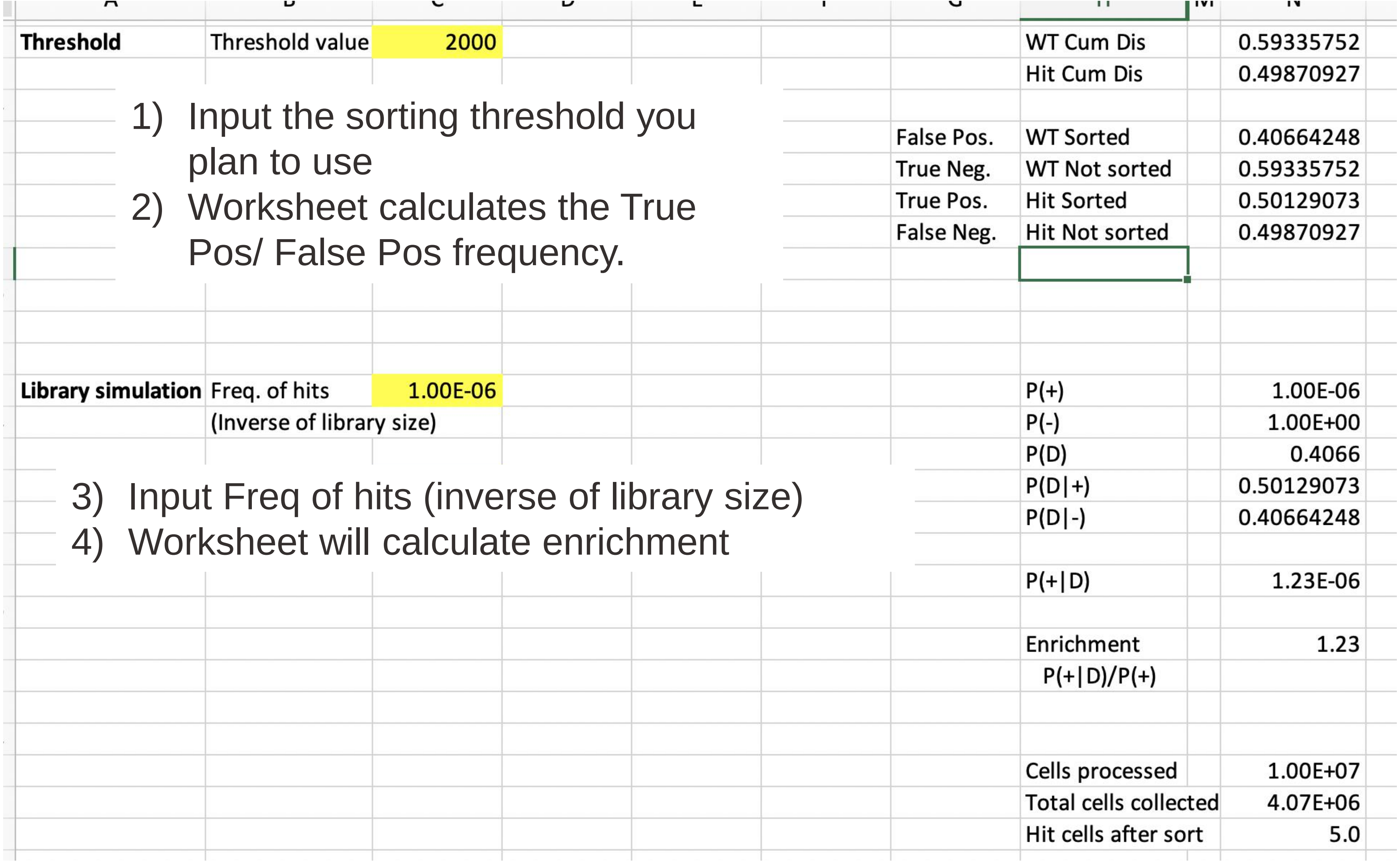
1) Use arithmetic mean worksheet (for now)
2) Input values in yellow blocks for WT and hit
3) Worksheet will autoplot on log and linear scale.

				CALCULATED			x	f(x)
WT / unimproved	Arith. Mean	2000		mu	7.489331		270.3744	1.0478E-06
	Arith. Std. Dev.	1000		sigma	0.472381		326.6078	3.9661E-06
							394.5368	1.2792E-05
							476.5938	3.5159E-05
							575.7174	8.2346E-05
							695.457	0.00016435
							0.1005	0.00027951
							14.827	0.00040507
							25.895	0.00050025
							480.86	0.00052645
							88.854	0.00047211
							13.089	0.00040661
							65.433	0.00032899
							49.407	0.00025005
							68.979	0.00017854
							28.609	0.00011976
							33.319	7.5463E-05
							088.76	4.467E-05
							01.291	2.484E-05
							78.068	1.2976E-05
							27.145	6.3678E-06

				CALCULATED			x	f(x)
Hit	Arith. Mean	2200		mu	7.602305		353.8222	8.7278E-07
	Arith. Std. Dev.	1000		sigma	0.433378		420.7949	3.3554E-06
							500.4444	1.0993E-05
							595.1703	3.0688E-05
							707.8262	7.3004E-05
							841.806	0.00014799
							1001.146	0.00025565
							1190.646	0.00037633
							1416.016	0.00047206
							1684.044	0.0005046
							2002.806	0.00045963
							2231.99	0.00039974
							2487.4	0.0003266
							2772.037	0.00025067
							3089.245	0.00018074
							3442.752	0.00012242
							3836.711	7.7894E-05
							4275.752	4.6561E-05
							4765.033	2.6145E-05
							5310.302	1.3792E-05
							5917.968	6.8344E-06
				CALCULATED				

https://bit.ly/2SnRoEf

Sorting calculations



Threshold	Threshold value	2000		WT Cum Dis	0.59335752
				Hit Cum Dis	0.49870927
			False Pos.	WT Sorted	0.40664248
			True Neg.	WT Not sorted	0.59335752
			True Pos.	Hit Sorted	0.50129073
			False Neg.	Hit Not sorted	0.49870927
Library simulation	Freq. of hits	1.00E-06		P(+)	1.00E-06
	(Inverse of library size)			P(-)	1.00E+00
				P(D)	0.4066
				P(D|+)	0.50129073
				P(D|-)	0.40664248
				P(+|D)	1.23E-06
				Enrichment	1.23
				P(+|D)/P(+)	
				Cells processed	1.00E+07
				Total cells collected	4.07E+06
				Hit cells after sort	5.0

1) Input the sorting threshold you plan to use
2) Worksheet calculates the True Pos/ False Pos frequency.

3) Input Freq of hits (inverse of library size)
4) Worksheet will calculate enrichment

Things to try

https://bit.ly/2SnRoEf

Set

	Mean	**Std. Dev.**
WT	2000	500
Hit	2200	500

Library = 10^6

What parameter defines best? **Enrichment**

What is the enrichment at threshold of 3200? **Enrichment = 1.77**

What percent of hits are ***not*** sorted at this threshold? **96%**

How many rounds of enrichment to get to 20% hits?

n = 21.3 (or practically 22)

https://bit.ly/2SnRoEf

How many cells should I sort? (Oversampling)

In the last example, 96% of hits were not sorted.

So you would need to screen 100 hit cells for your screen to recover four of them.

If your library is 10^6 (as in last example)

How many hit cells will you sort if you sort

10^6 cells (1x)

10^7 cells (10x)

10^8 cells (100x)

https://bit.ly/2SnRoEf

Probability of missing a hit

Probability of missing 1 hit = False Neg. Rate

Probability of missing n hits = FN^n

(from previous example)

10^6 (1x)	$0.96^1 = 0.96$	
10^7 (10x)	$0.96^{10} = 0.66$	
10^8 (100x)	$0.96^{100} = 0.01$	(i.e. 99% chance of recovering at least one hit cell)

Other sources of failure in FACS

- Cell viability
  - Even though you sort 100 cells, how many of them are viable?
    - Specifically how many hits?

- Biases during growth
  - Between rounds of sorting, regrowing your library will allow fast growers to dominate the population

- Two-state sorting (biosensors)
  - Constitutive ‘on’ and constitutive ‘off’

    Forthcoming publication on this

Summary - High throughput screening

- The problem: How to interpret a negative screening result?

- Having a positive control is good.
  - If not, make some assumptions

- Estimate enrichment to inform the rounds of sorting you need

- Ensure you sort enough cells to confidently recover a hit

DIVERSIFICATION

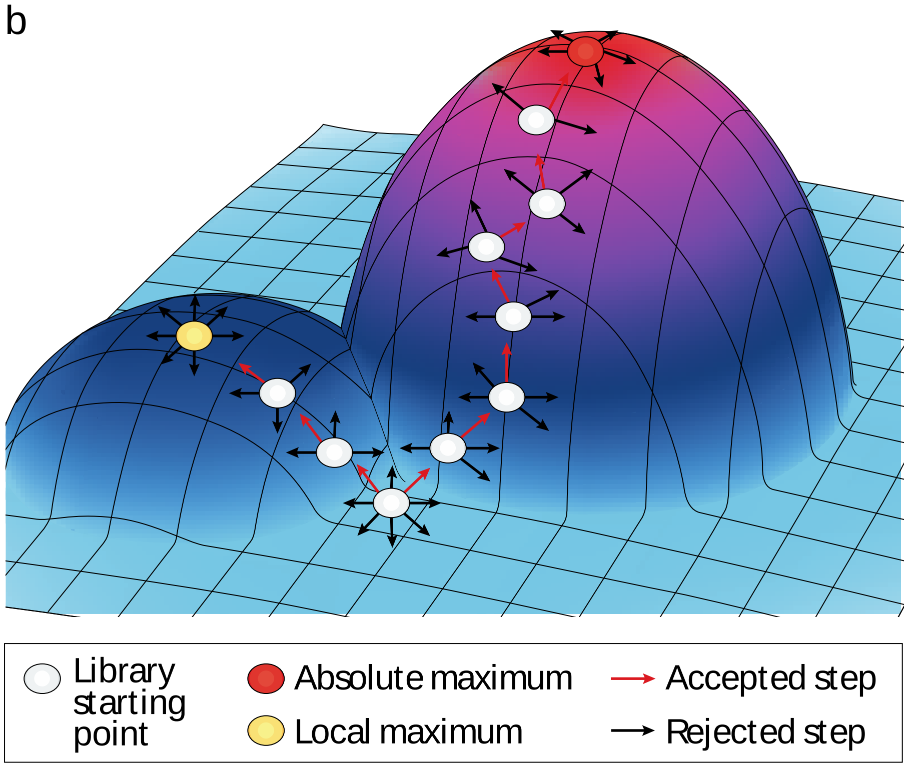
How should you generate mutant libraries?

Many methods exist

Error-prone PCR is common

(but has problems)

Sequence space is much larger than what is searchable



Packer, M. S., & Liu, D. R. (2015). *Nature Reviews. Genetics*.

Error-prone PCR

Pro's

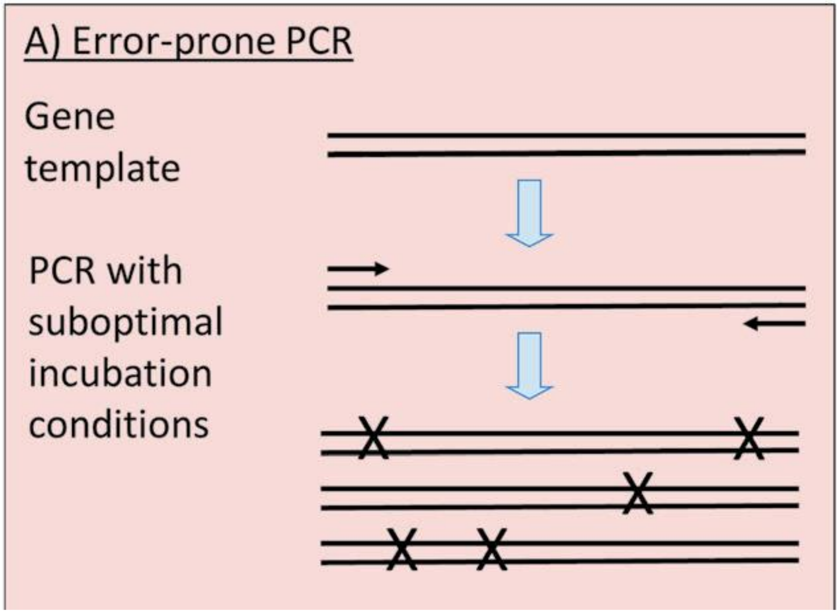
- Easy
- Requires no knowledge

Con's

- Bias for particular mutations
- Limited diversity

Given a codon, only feasible to access ~8 other codons

*GTC: **X**TC, G**X**C, GT**X***



Curran, A., Swainston, N., Day, P. J., & Kell, D. B. (2015). *Chemical Society Reviews*.

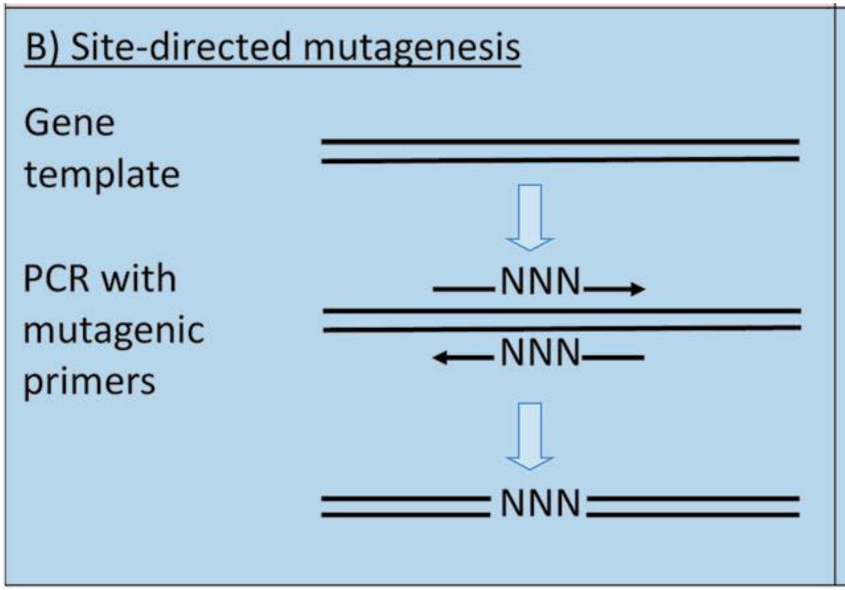
Oligo-based site directed mutagenesis

Pro's

- Focus diversity to particular parts of a protein
- Control the specific mutation result
  - NNK
  - Can access all 20 or particular subsets
  - Amenable to precise library designs

Con's

- More complicated



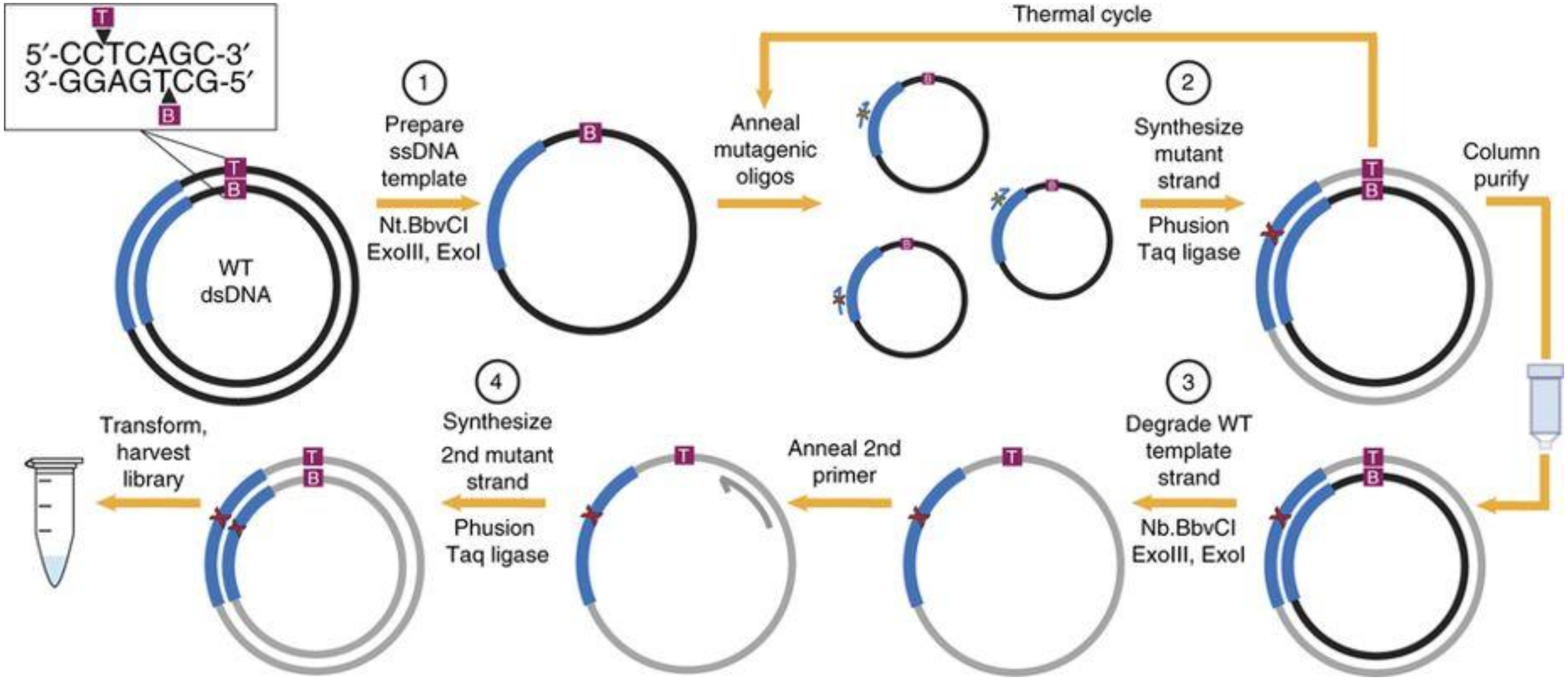
Codon compression algorithms for saturation mutagenesis

Pines, G., Pines, A., Garst, A. D., Zeitoun, R. I., Lynch, S. A., & Gill, R. T. (2015). *ACS Synthetic Biology*.

Curran, A., Swainston, N., Day, P. J., & Kell, D. B. (2015). *Chemical Society Reviews*.

Nicking mutagenesis

One day, in vitro



Wrenbeck, E. E., Klesmith, J. R., Stapleton, J. A., Adeniran, A., Tyo, K. E. J., & Whitehead, T. A. (2016). Plasmid-based one-pot saturation mutagenesis. *Nature Methods*, *13*(11), 928–930.

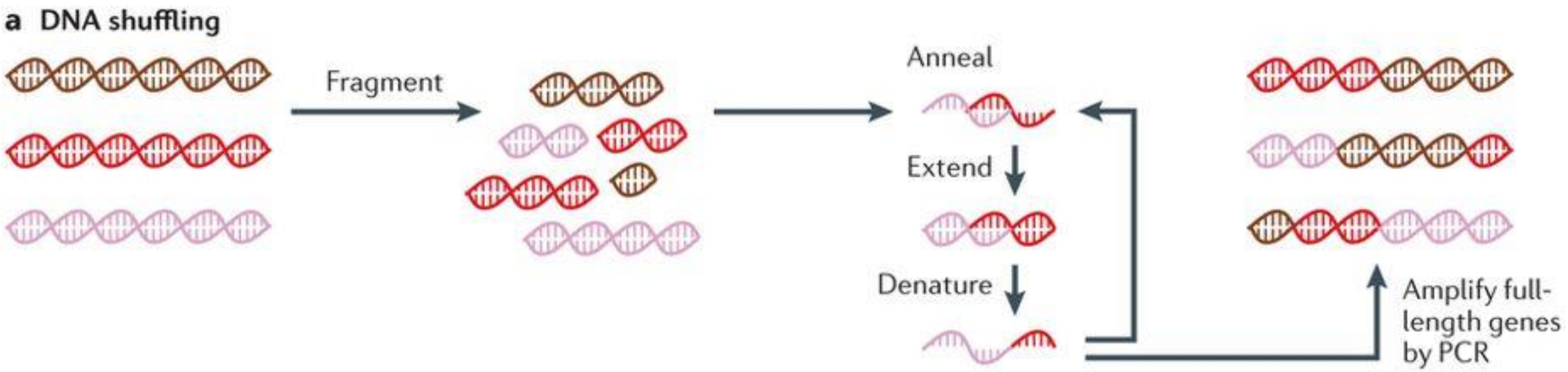
Shuffling

Pro’s

- Recombine natural diversity
- Recombine mutations from different hits

Con’s

- More complicated



Packer, M. S., & Liu, D. R. (2015). *Nature Reviews. Genetics*.

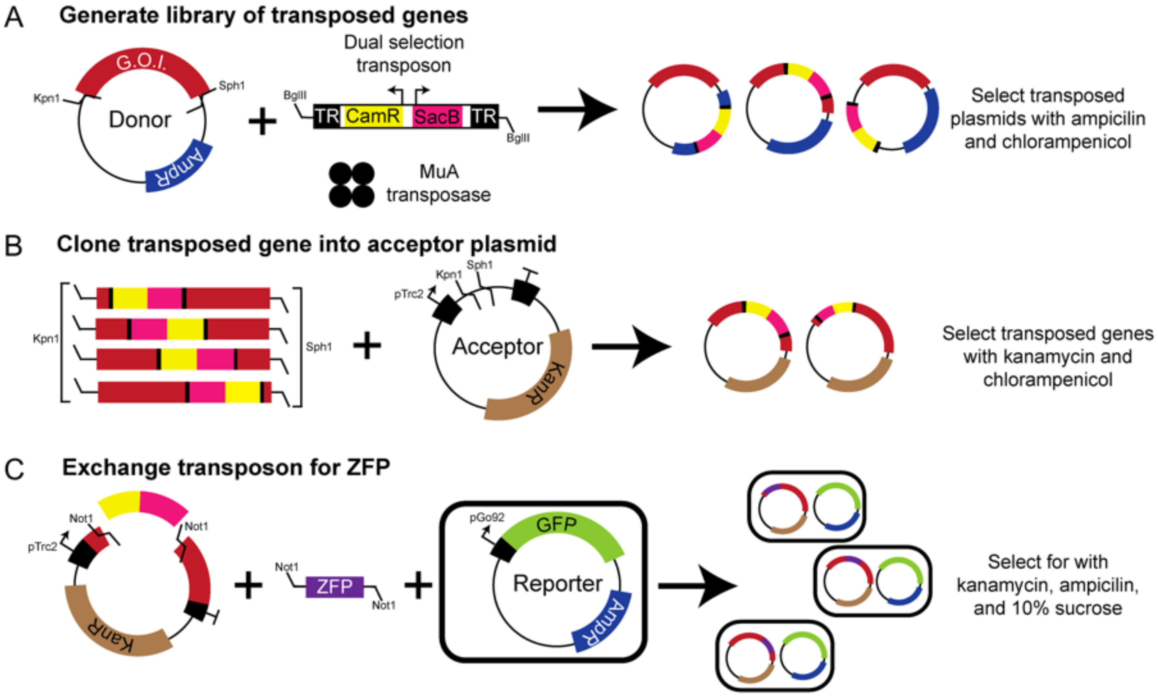
Domain insertion mutagenesis

Pro's

- Randomly combine two proteins/domains
- Engineering allostery

Con's

- Transposon mutagenesis is biased
- More complicated


A Generate library of transposed genes
G.O.I.
Donor
Kpn1
Sph1
AmpR
+
Dual selection transposon
BglII
TR CamR SacB TR
BglII
MuA transposase
Select transposed plasmids with ampicilin and chlorampenicol
B Clone transposed gene into acceptor plasmid
Kpn1
Sph1
+
pTrc2
Kpn1
Sph1
Acceptor
KanR
Select transposed genes with kanamycin and chloramphenicol
C Exchange transposon for ZFP
Not1
Not1
pTrc2
KanR
+
Not1
ZFP
Not1
+
pGo92
GFP
Reporter
AmpR
Select for with kanamycin, ampicilin, and 10% sucrose

Younger, A. K. D., … Tyo, K. E. J., & Leonard, J. N. (2018). Development of novel metabolite-responsive transcription factors via transposon-mediated protein fusion. *Protein Engineering, Design and Selection*, *31*(2), 55–63.

Nadler, D. C., …., & Savage, D. F. (2016). Rapid construction of metabolite biosensors using domain-insertion profiling. *Nature Communications*.

Summary

- Many ways to create diversity that
  - reduce library size
  - focusing on mutations with high(er) probability of success

- Things I didn't talk about
  - Using Illumina sequencing to characterize libraries
    - Naïve library
    - Sorted library

  - Using bioinformatics and structural information to select strategies for rational library design

Conclusions

- Directed evolution is a powerful tool for engineering biology.

- High throughput screens
  - Single cell measurements can be very noisy
  - Statistical approaches can ensure a robust screening strategy

- Diversification
  - Sequence space is large
  - Focused/rational library approaches
    - reduce screening effort
    - focus on more likely candidates

Acknowledgements

Tyo lab

- Adebola Adeniran
- Dante Pertusi
- Sarah Stainbrook
- Jessica Yu

Feedback

https://forms.gle/9941g9uZGdaFcTAD6

Collaborators

- Josh Leonard (Northwestern)
- Tim Whitehead (Colorado)

Funding

- NSF
  - CAREER *CBET-1452549*
  - MCB-1614953
  - GOALI *CBET-1803747*
- DOE
  - BER DE-SC0019339
- NIH
  - NIMH R01 MH103910
  - NINDS UF1 NS10769701