

Why don't we agree? Studying influenza with RNA interference

Michael Newton
UW Madison



From a project on influenza biology with:

Qiuling He

Lin Hao

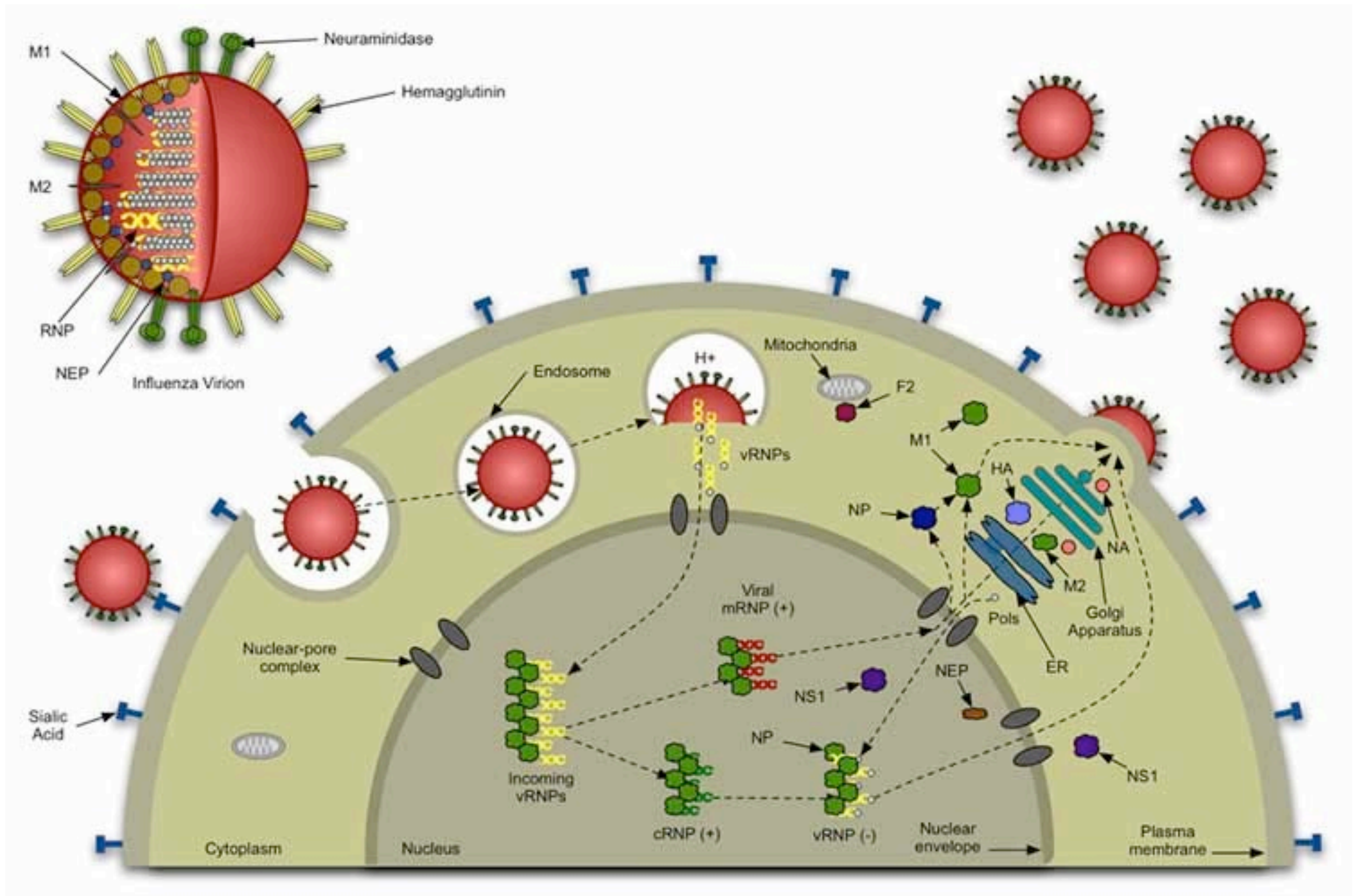
Mark Craven

Paul Ahlquist

Thanks Christina Kendzierski...

What human genes does influenza virus co-opt during its life cycle?

a bit about flu

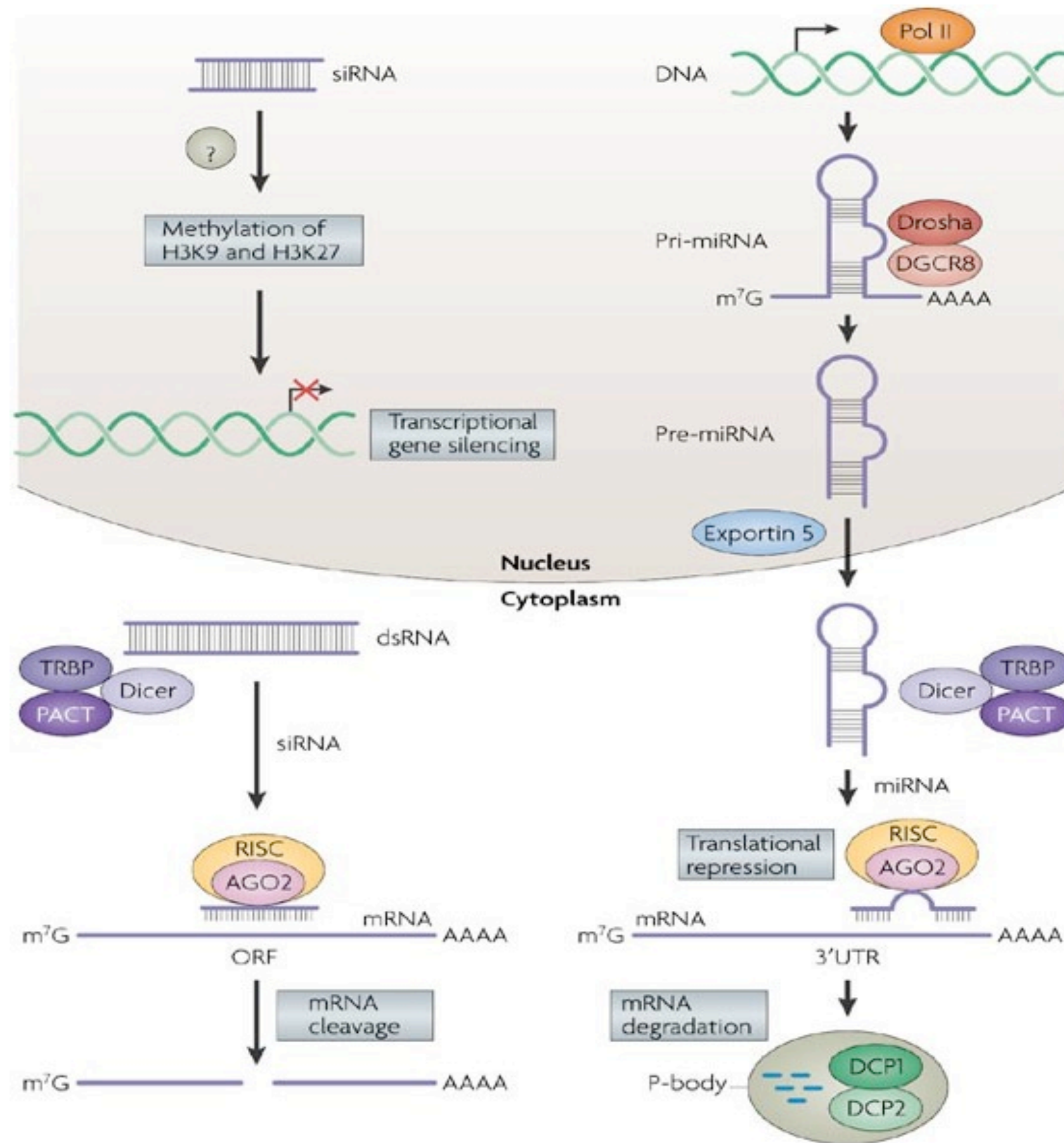


Experimental Method: RNAi

RNA interference is a process within living cells that moderates the activity of their genes.

Fire and Mello, 2006, Nobel Prize

a bit about RNAi



Nature Reviews | Genetics

Or, more simply, ...

RNA interference =



Or, more simply, ...

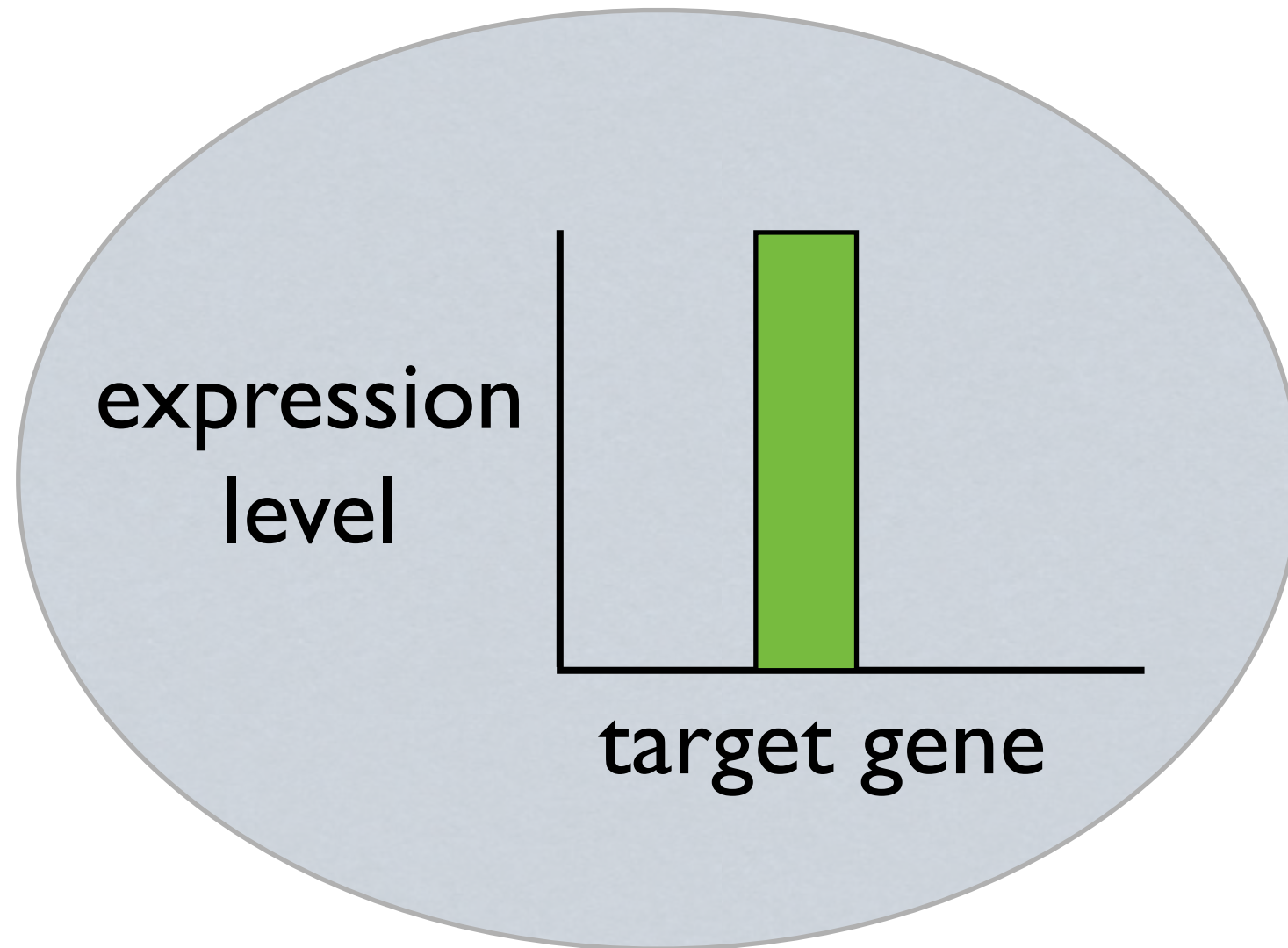
RNA interference =



genome-wide...

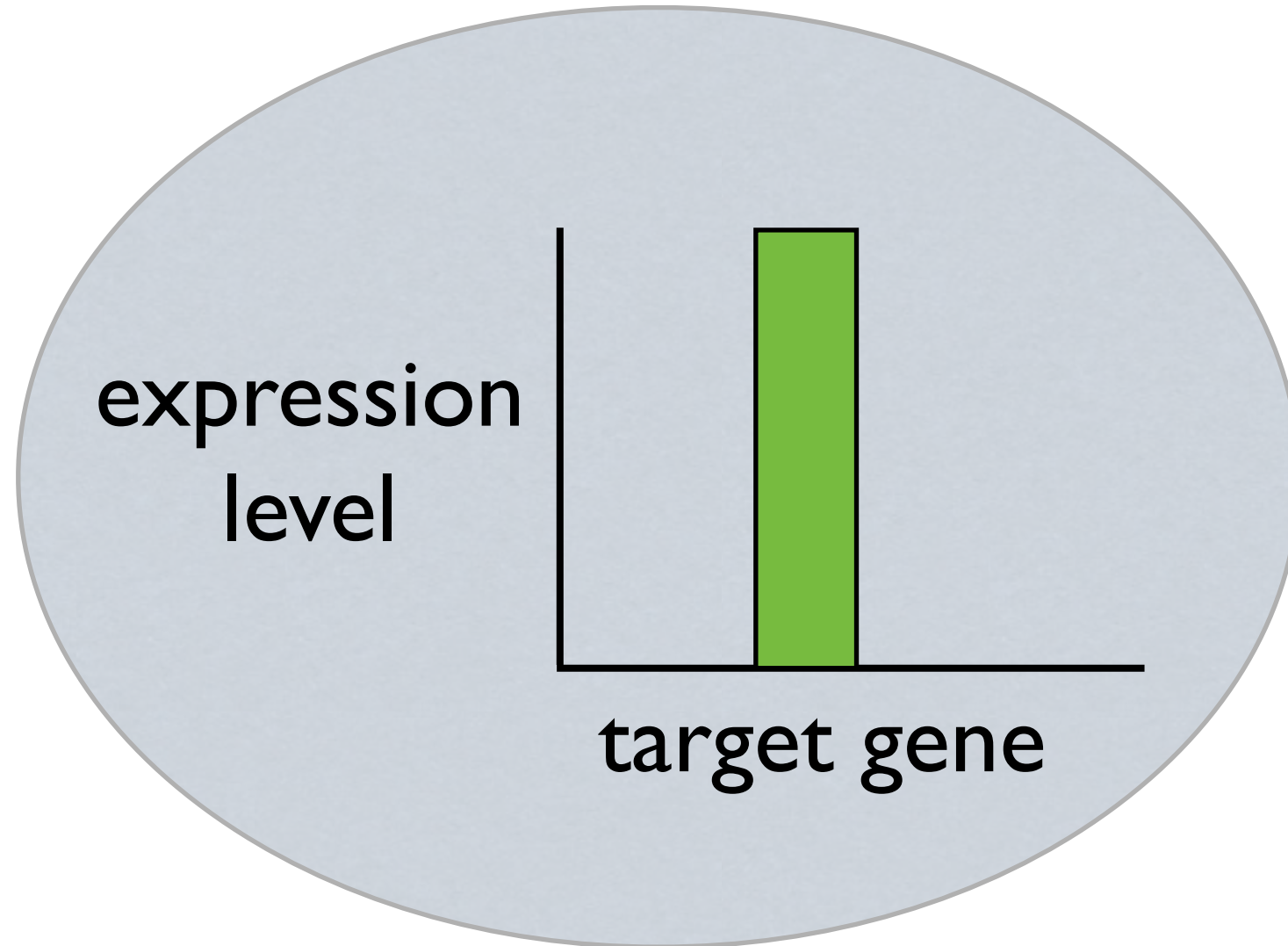
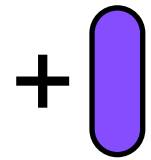


cell



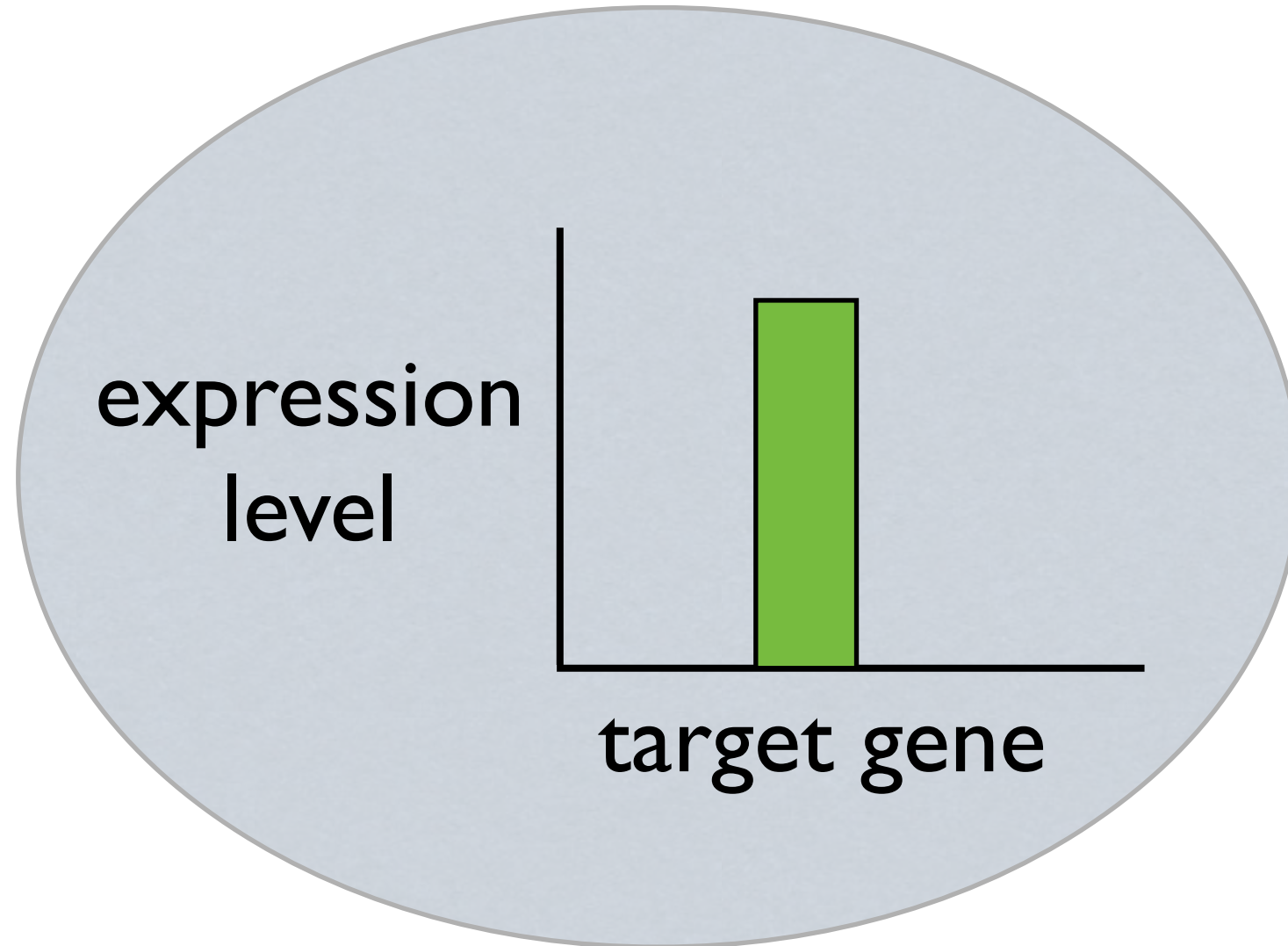
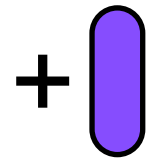
cell

add siRNA



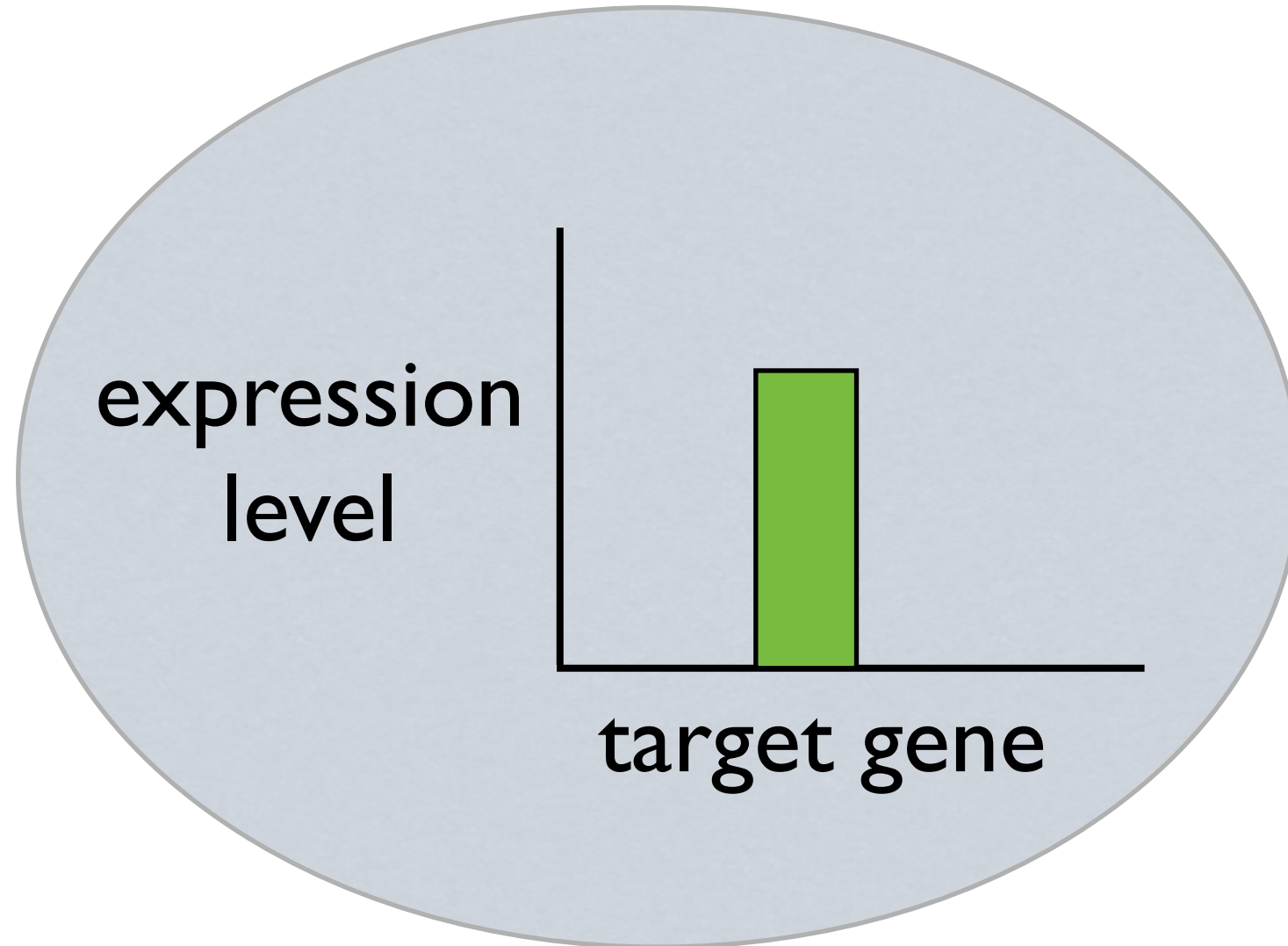
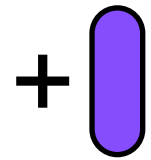
cell

add siRNA



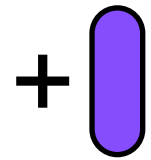
cell

add siRNA

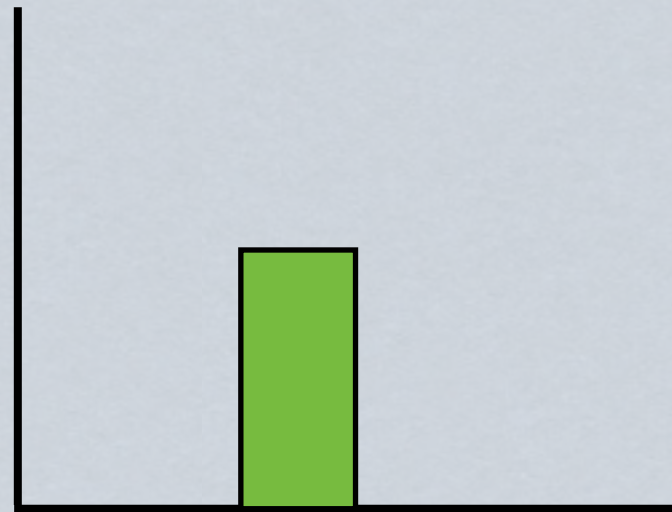


cell

add siRNA



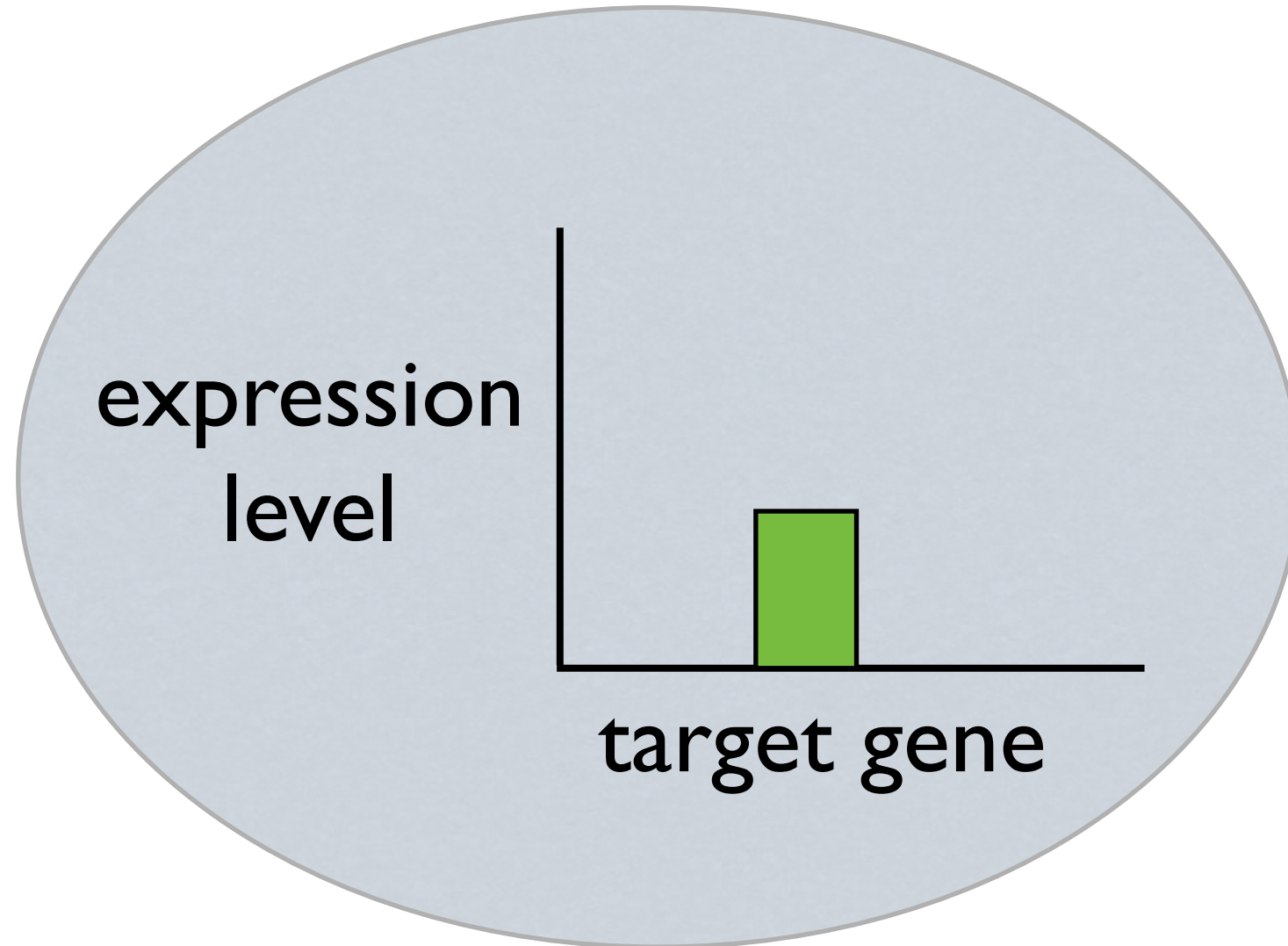
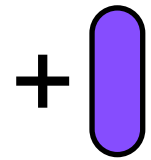
expression
level



target gene

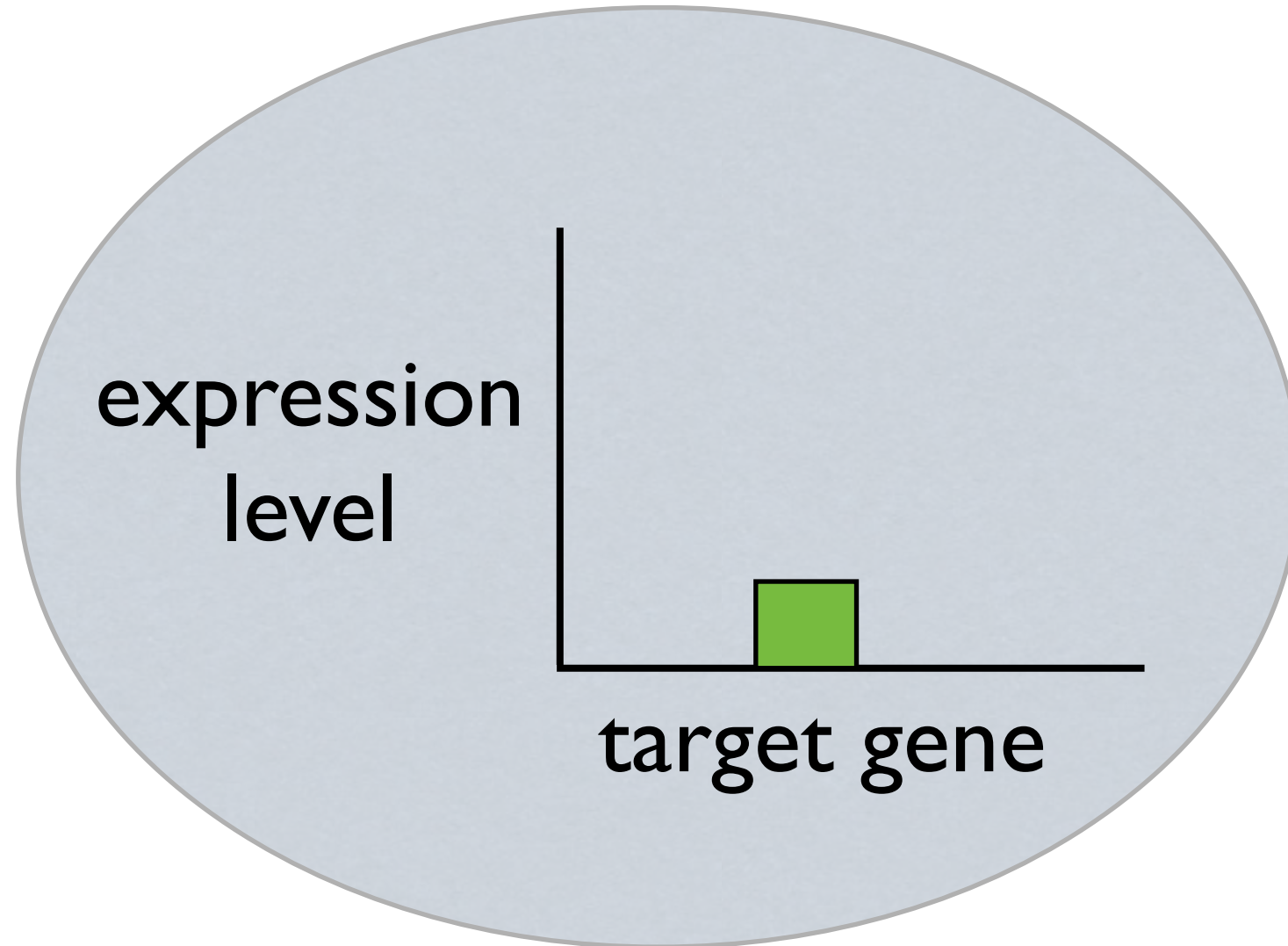
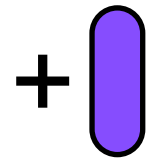
cell

add siRNA

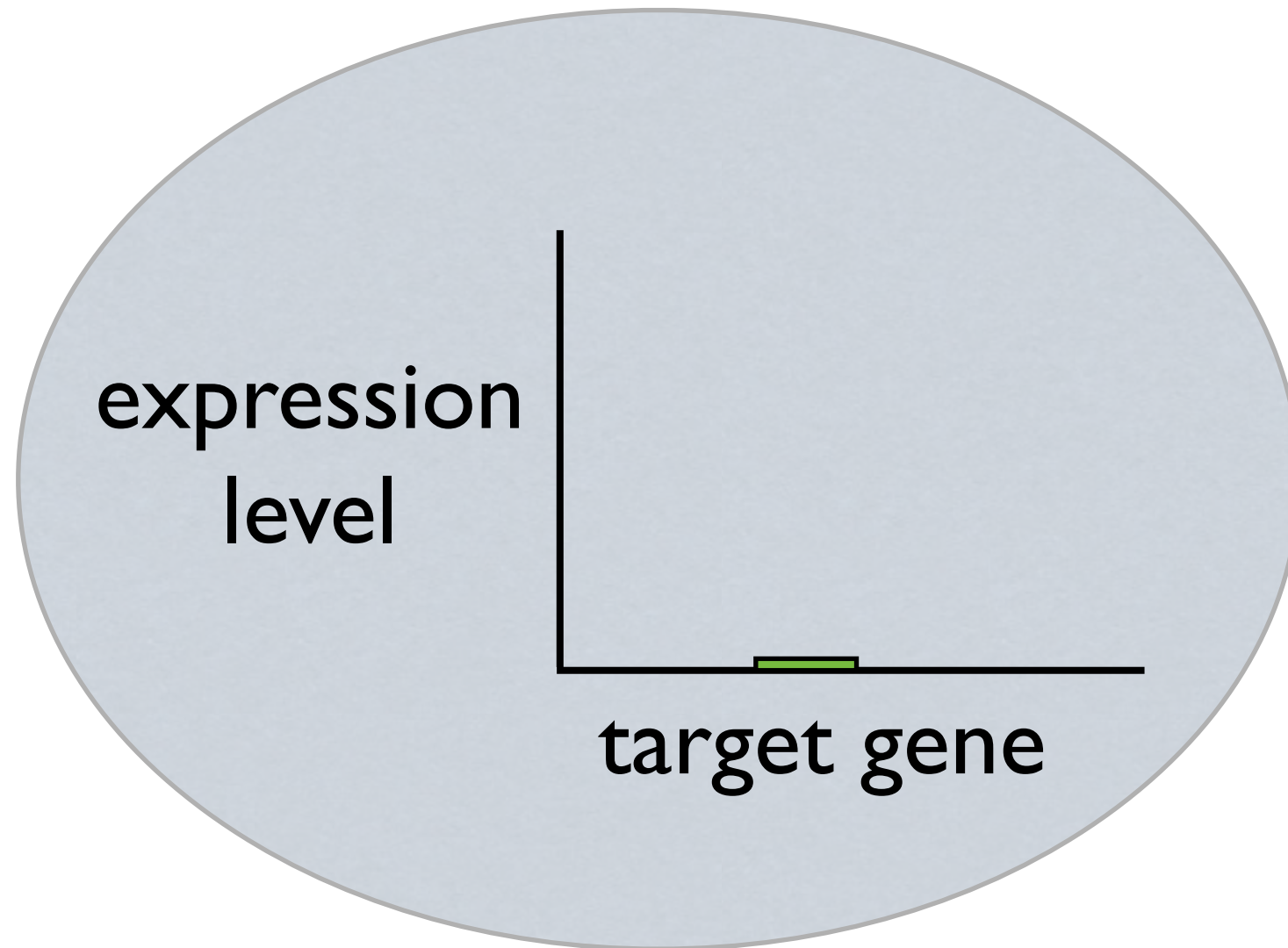


cell

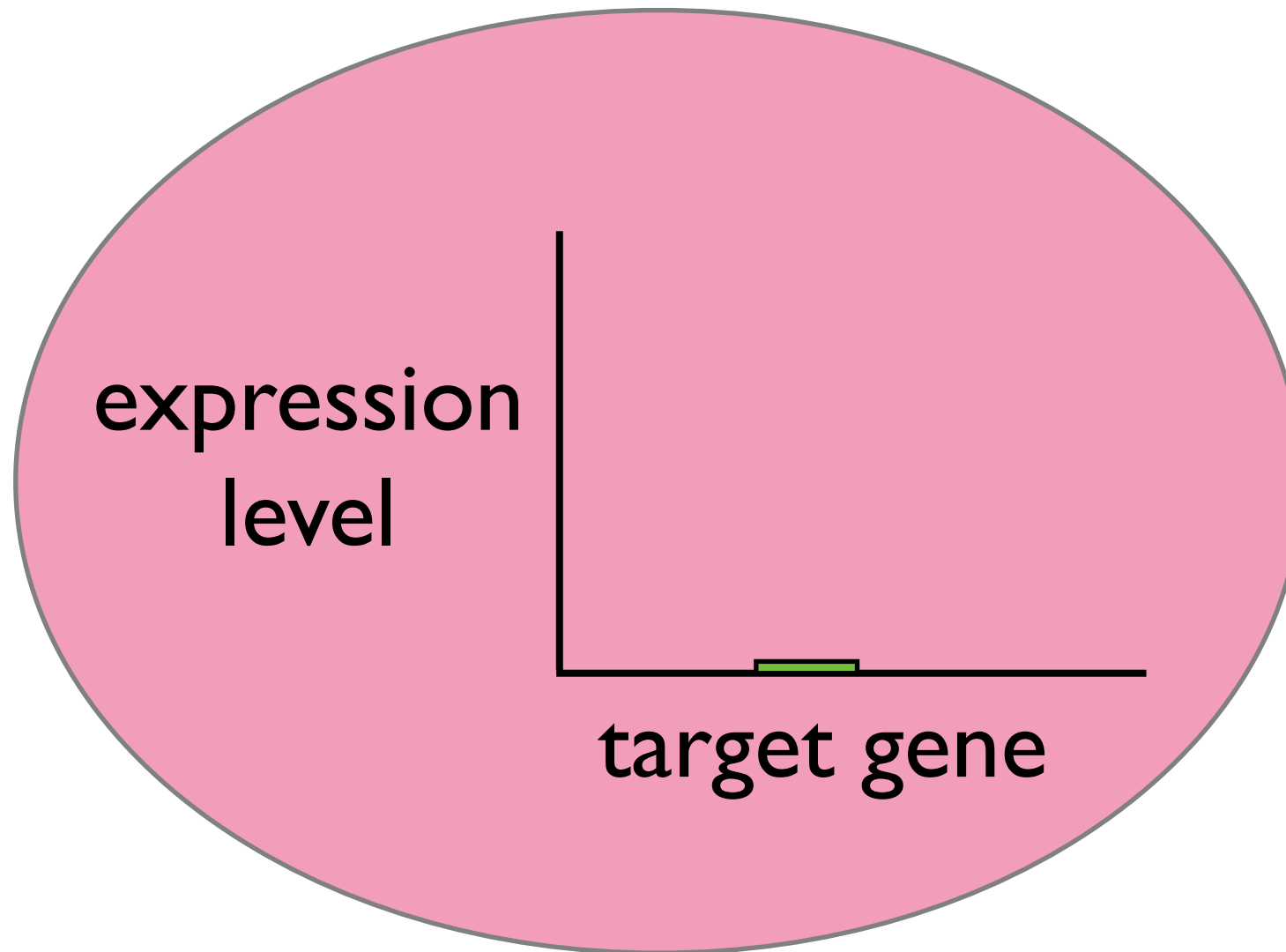
add siRNA



cell



cell

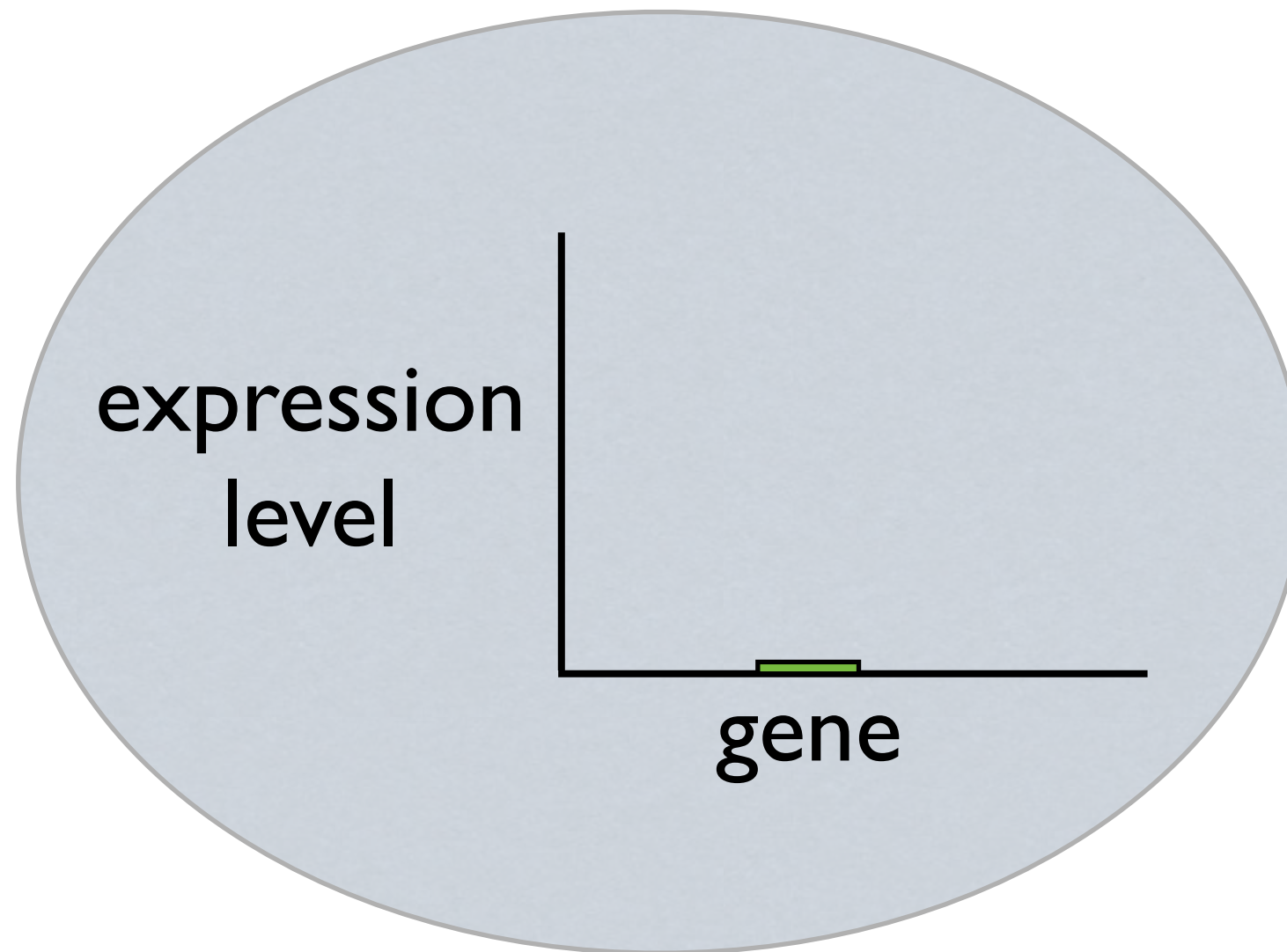


phenotype of interest changes

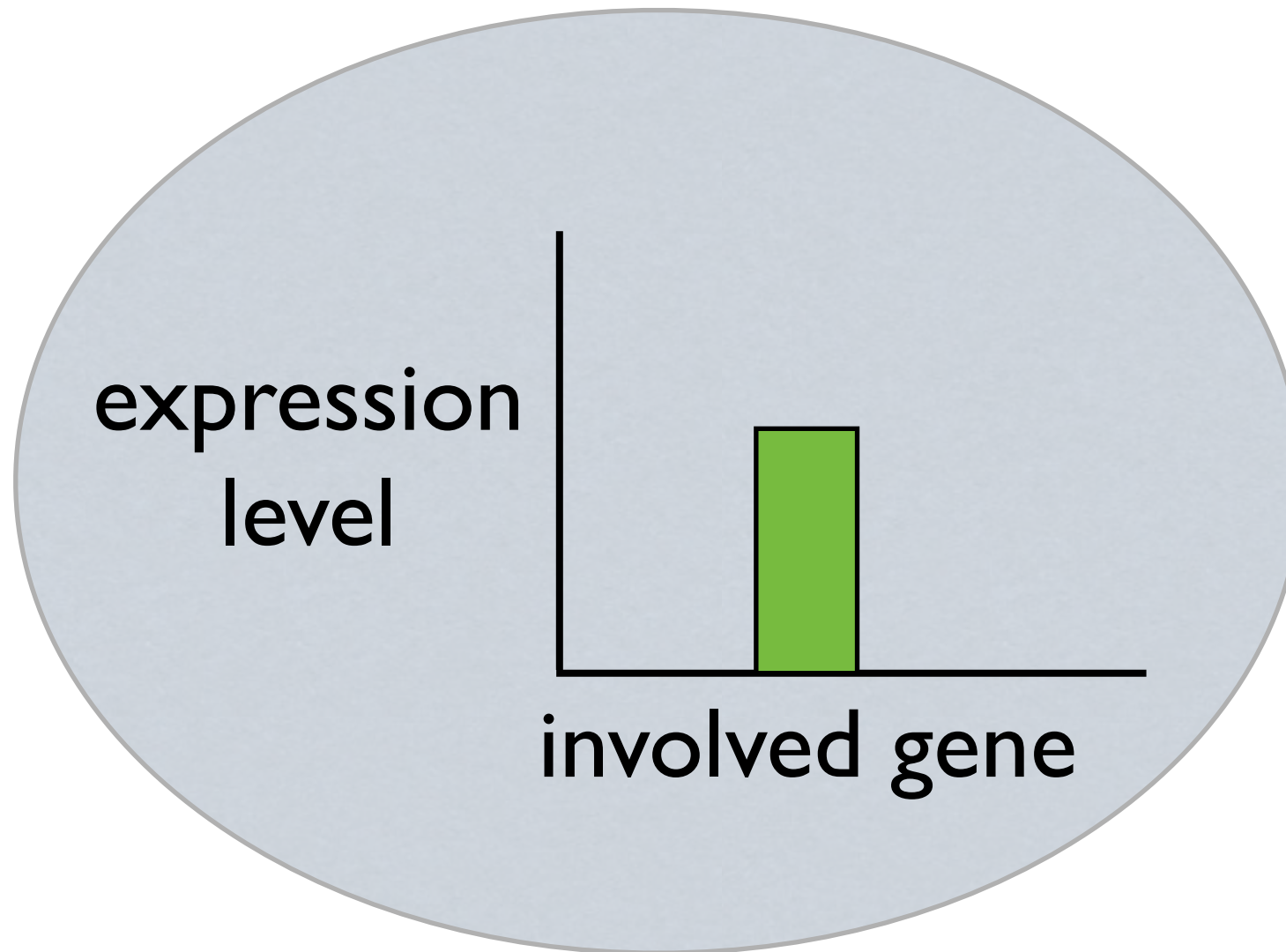
explain phenotype

Issues

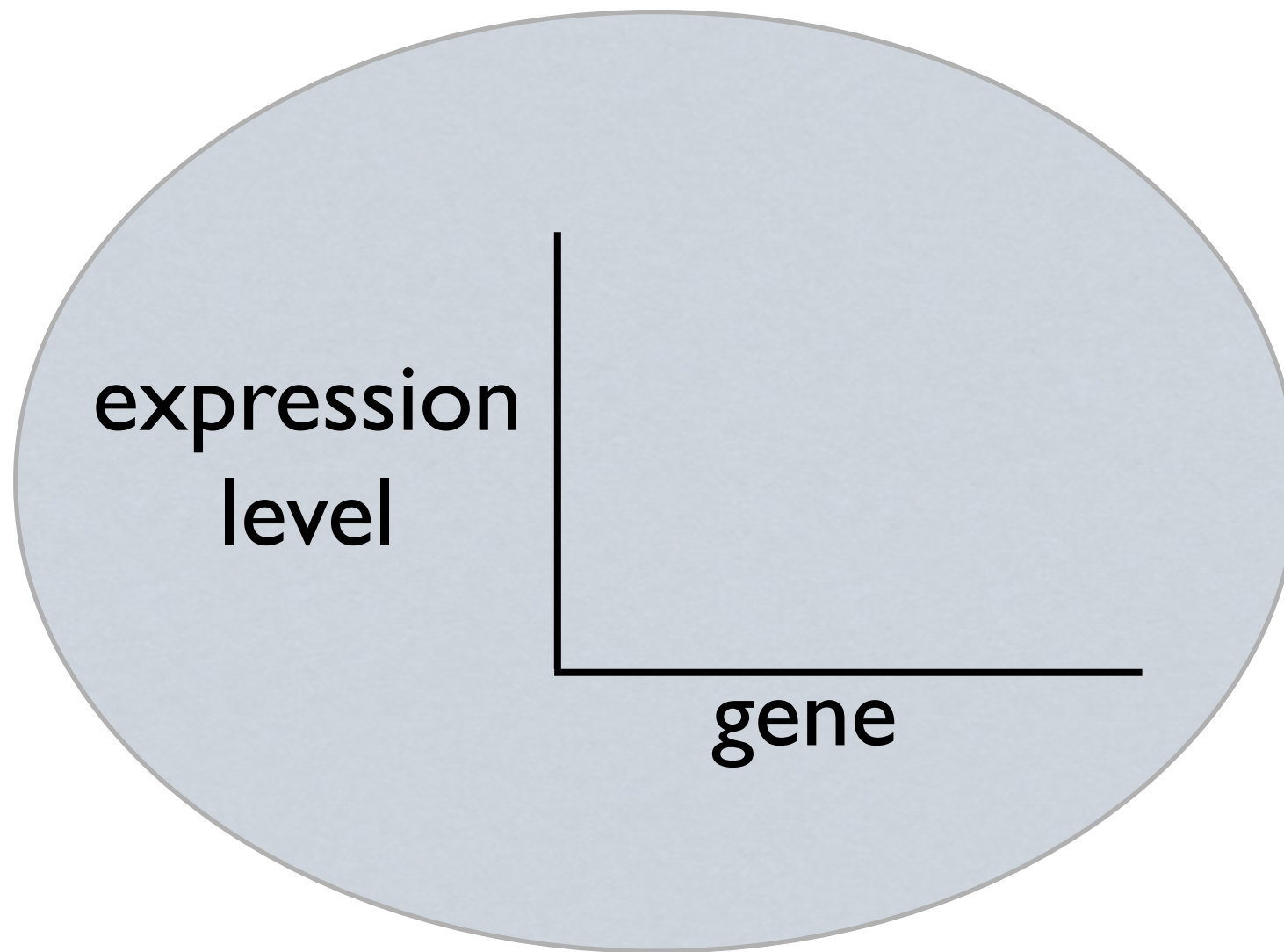
Involvement: gene may not affect phenotype



Efficiency: knockdown may not be complete

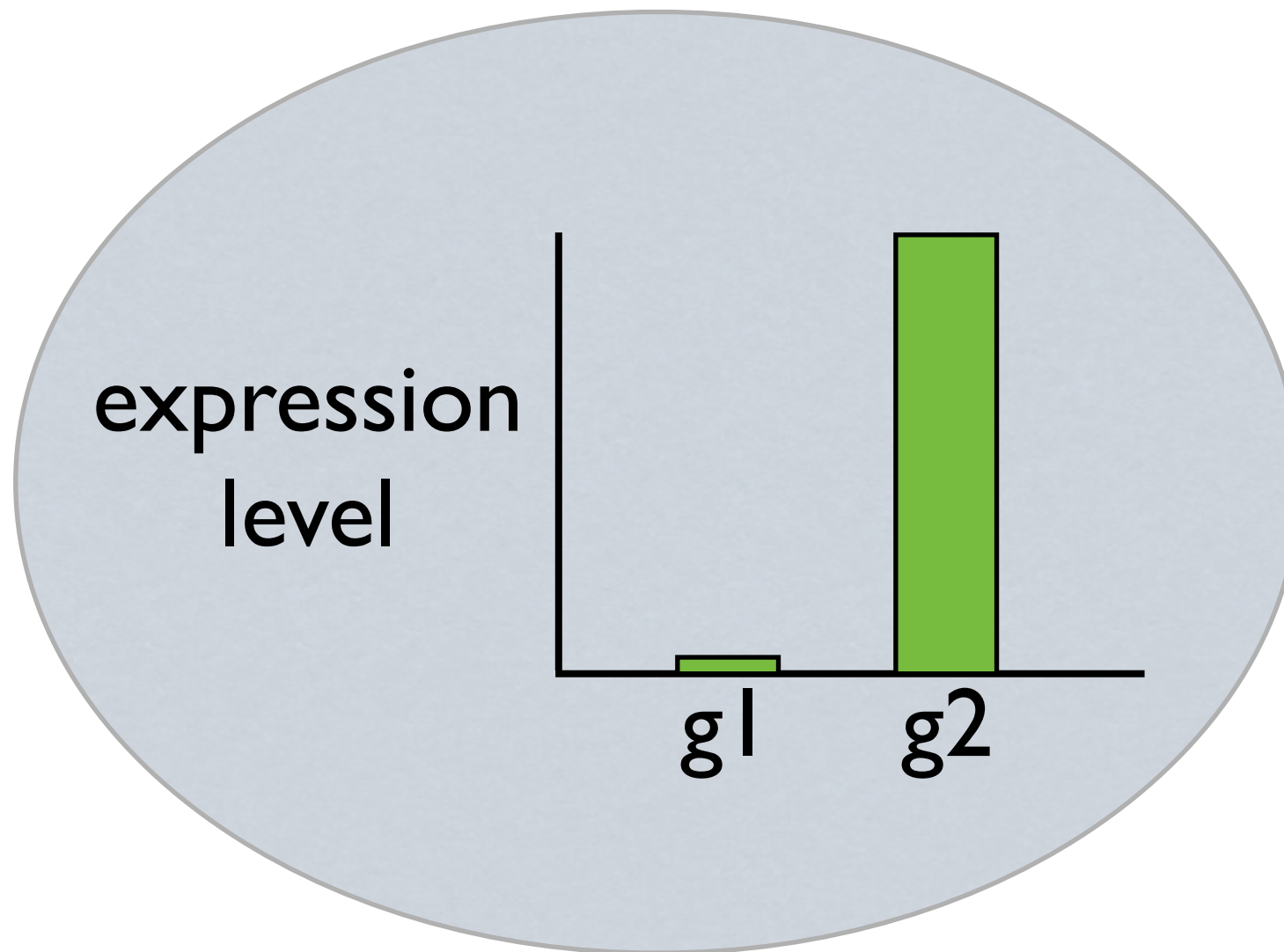


Accessibility: something blocks the phenotype



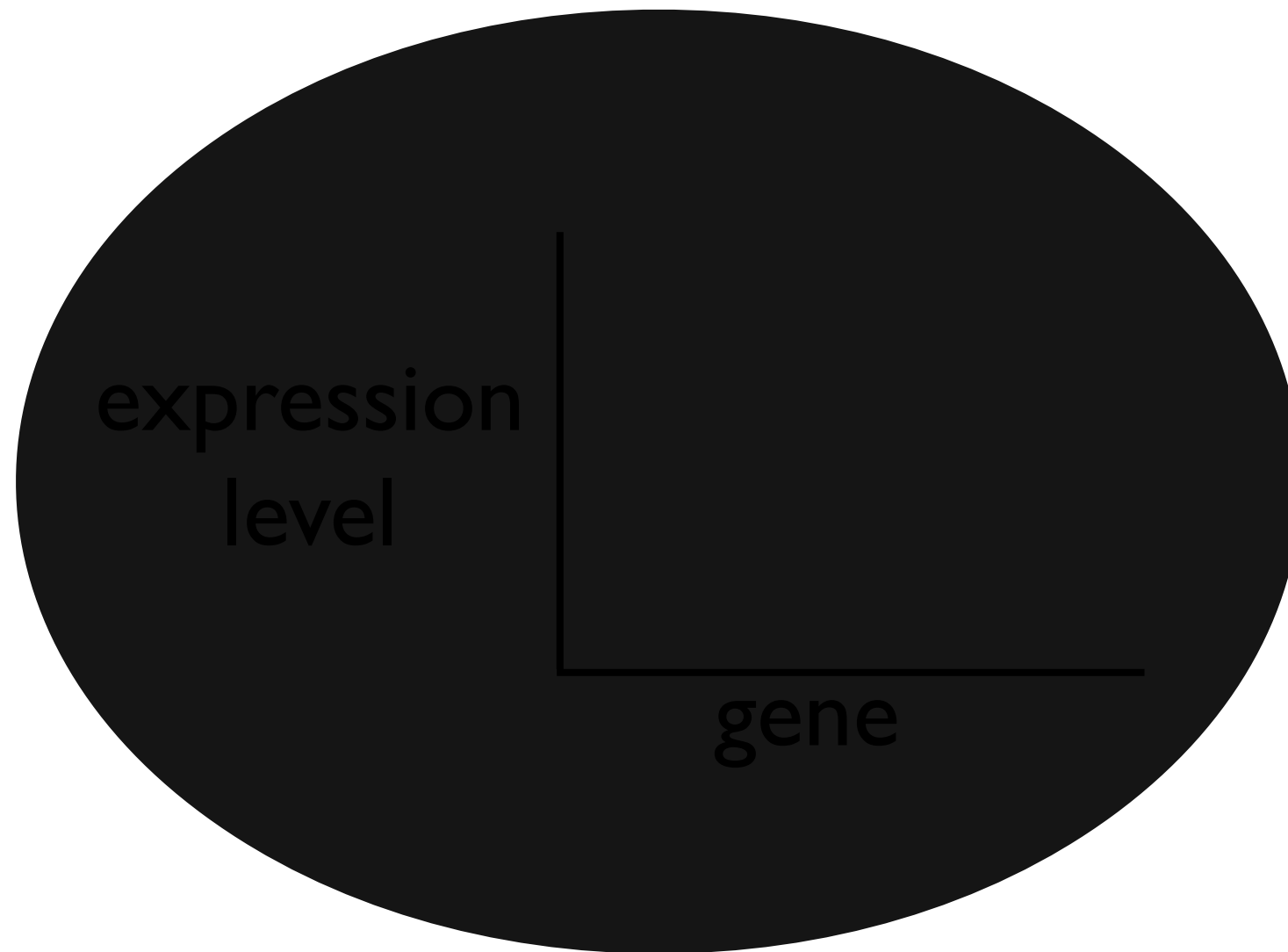
a. no expression in these particular cells

Accessibility: something blocks the phenotype



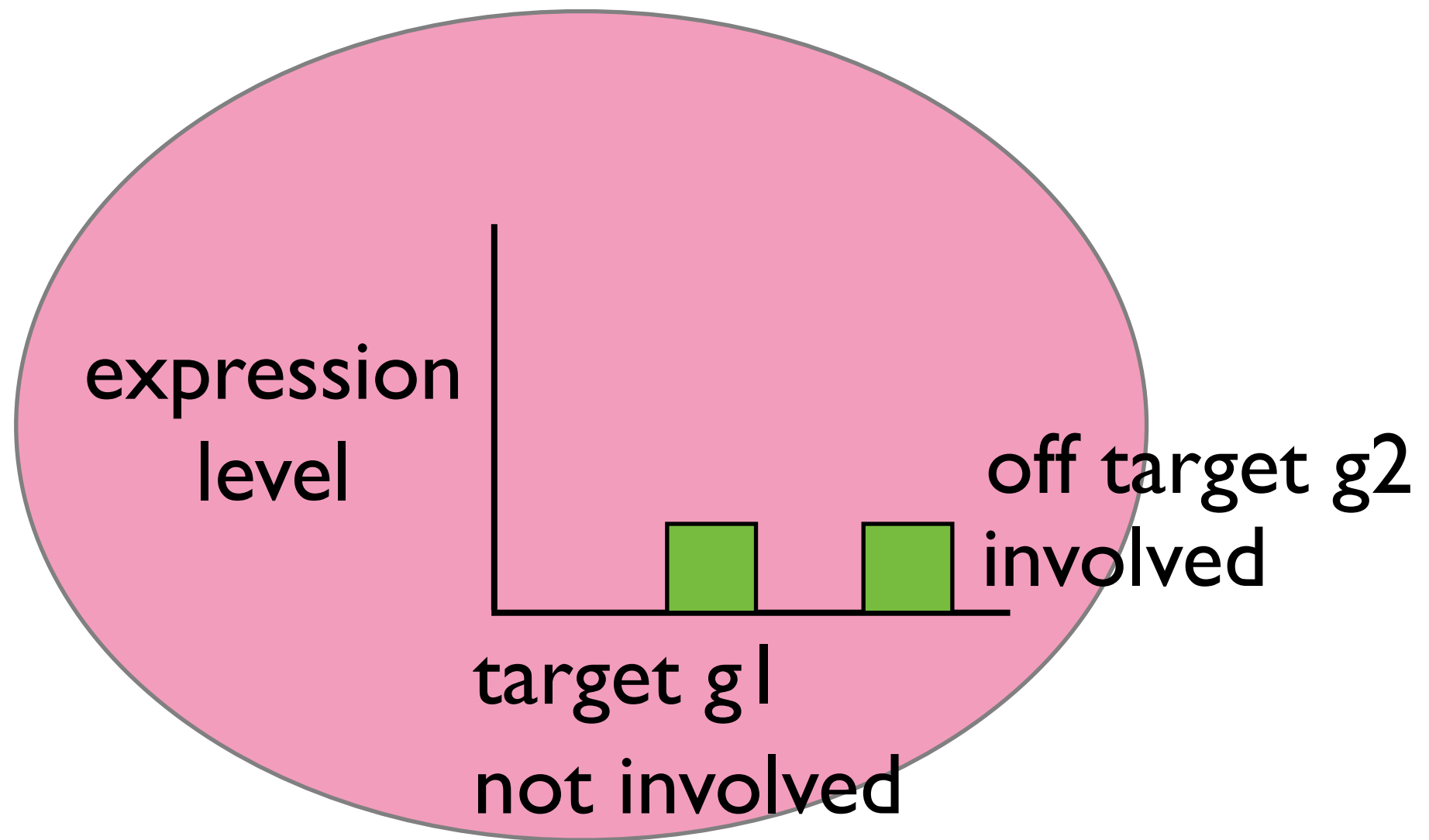
b. Redundancy/masking

Accessibility: something blocks the phenotype



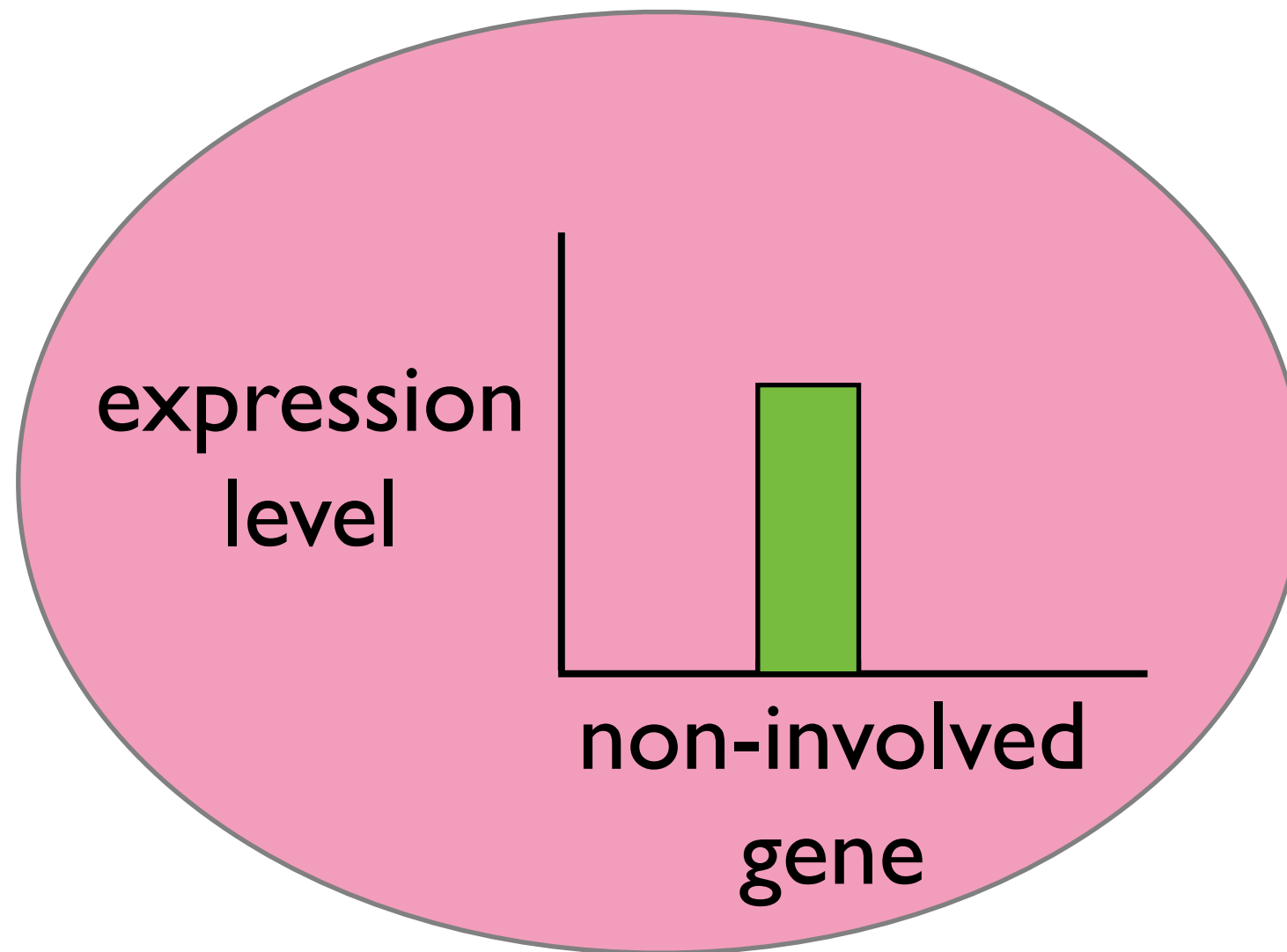
c.cytotoxicity

Off target effects



Measurement error

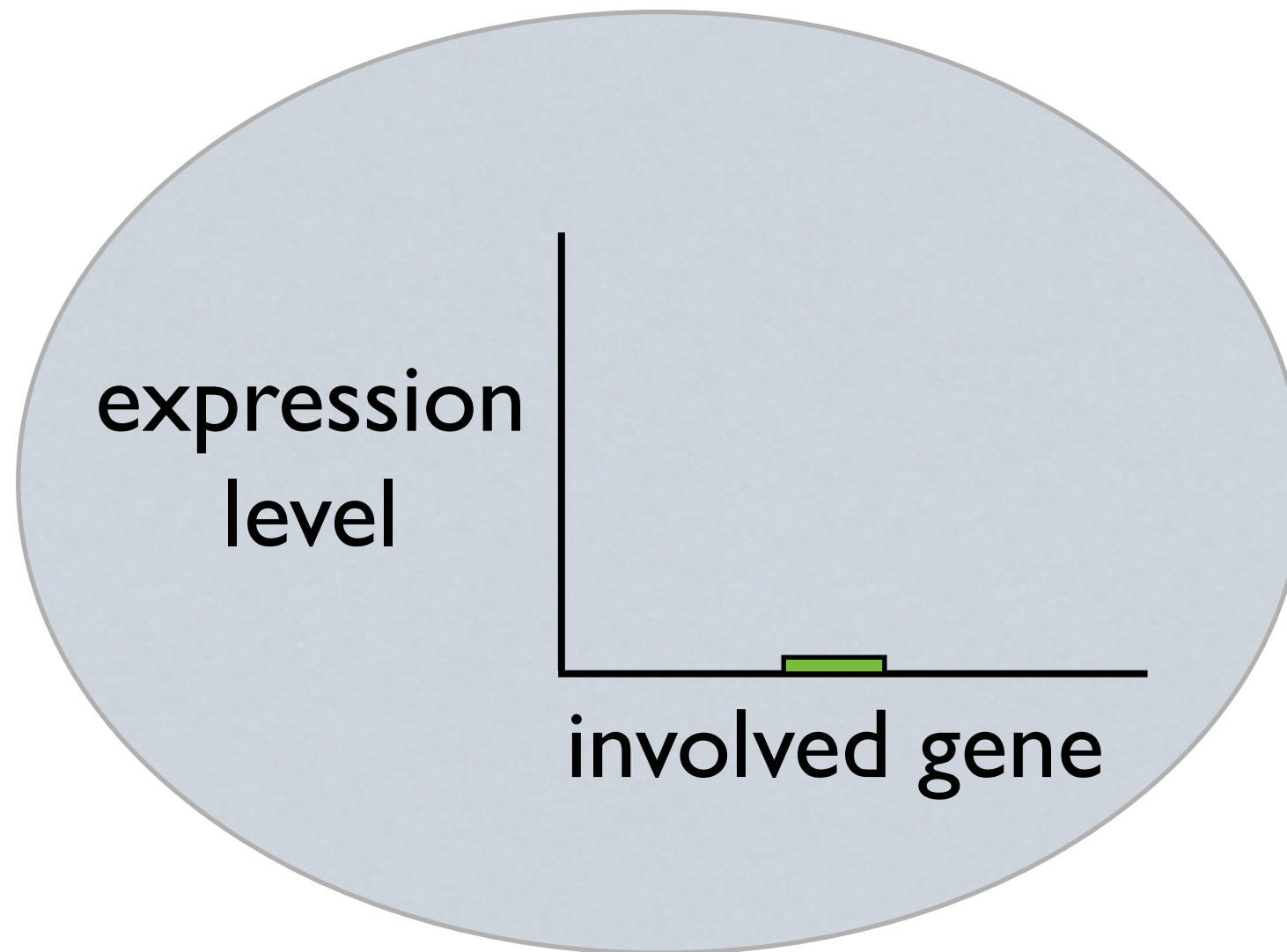
Measurement error



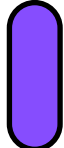
false positive

Measurement error

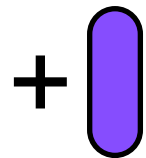
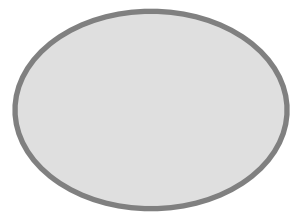
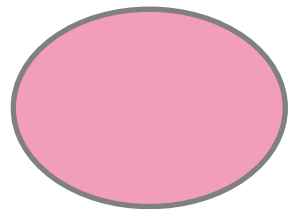
Measurement error



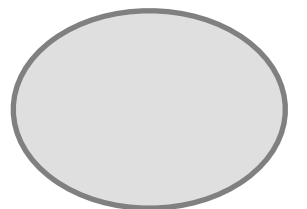
false negative

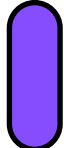
one siRNA 

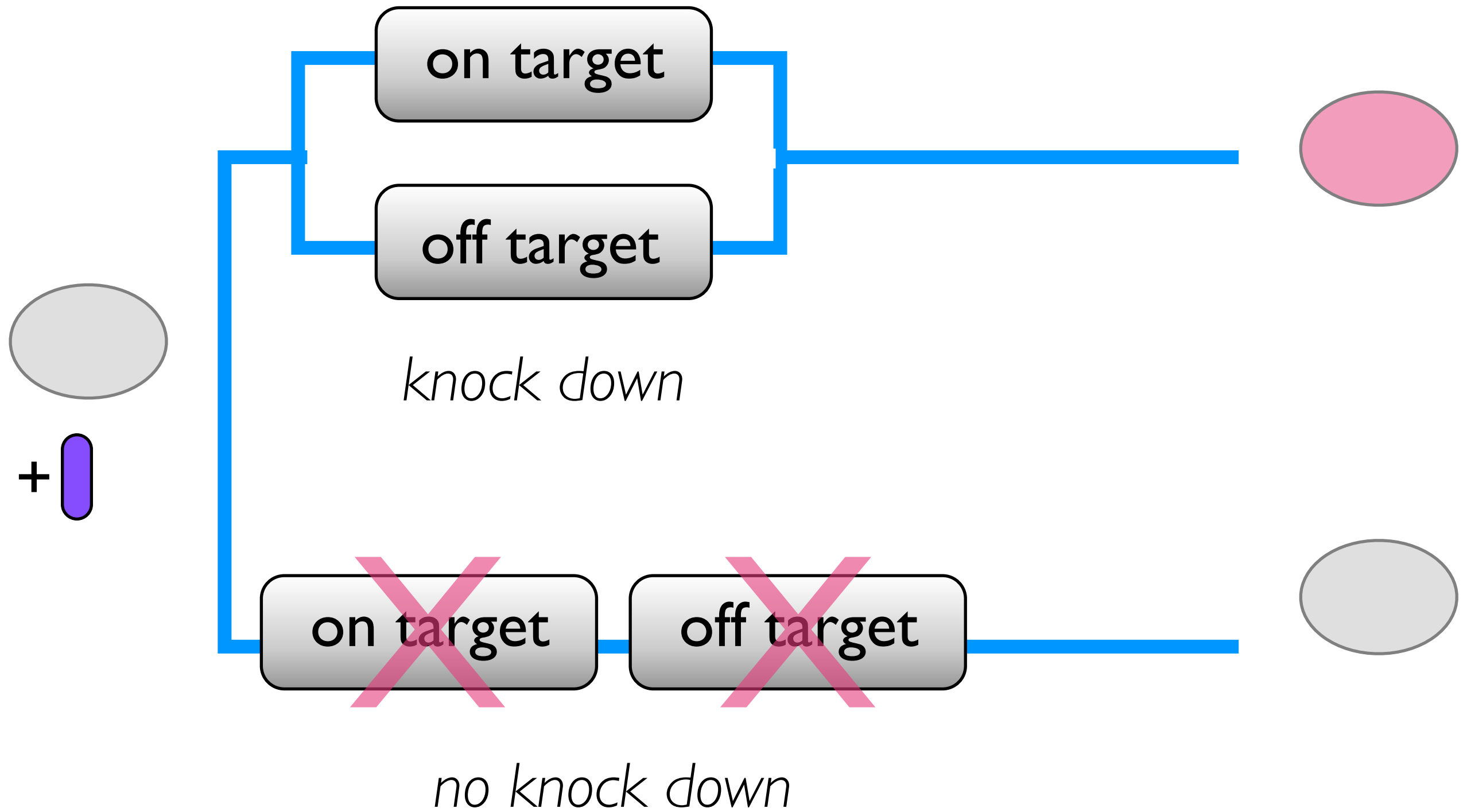
knock down

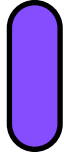


no knock down

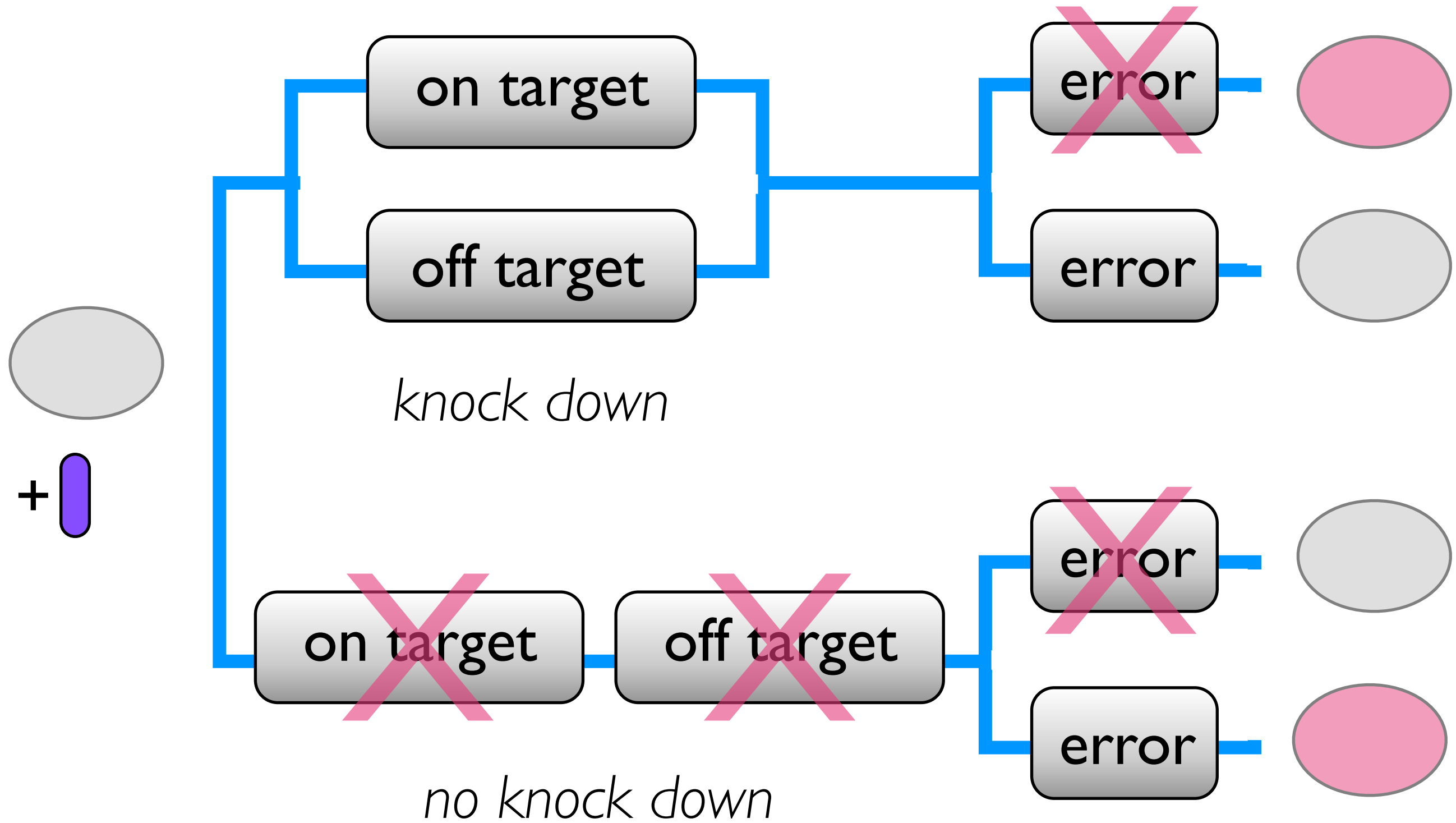


one siRNA 



one siRNA 

measurement

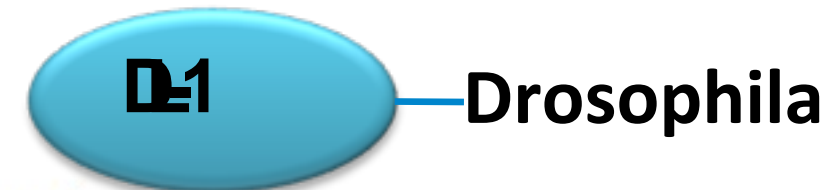


Meta analysis of four recent studies

nature
2008

***Drosophila* RNAi screen identifies host genes important for influenza virus replication**

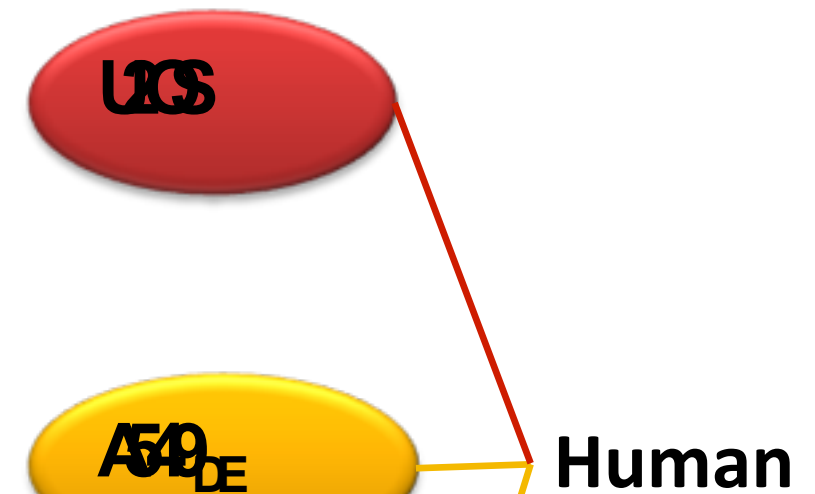
Linhui Hao^{1,2*}, Akira Sakurai^{3*†}, Tokiko Watanabe³, Ericka Sorensen¹, Chairul A. Nidom^{5,6}, Michael A. Newton⁴, Paul Ahlquist^{1,2} & Yoshihiro Kawaoka^{3,7,8,9}



Cell
2009

The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus

Abraham L. Brass,^{1,2,4,9,*} I-Chueh Huang,^{5,9} Yair Benita,^{3,10} Sinu P. John,^{1,10} Manoj N. Krishnan,⁶ Eric M. Feeley,¹ Bethany J. Ryan,¹ Jessica L. Weyer,⁵ Louise van der Weyden,⁸ Erol Fikrig,^{6,7} David J. Adams,⁸ Ramnik J. Xavier,^{2,3} Michael Farzan,^{5,*} and Stephen J. Elledge^{4,*}



nature
2010

Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication

Alexander Karlas^{1*}, Nikolaus Machuy^{1*}, Yujin Shin¹, Klaus-Peter Pleissner², Anita Artarini¹, Dagmar Heuer¹, Daniel Becker¹, Hany Khalil¹, Lesley A. Ogilvie¹, Simone Hess^{1†}, André P. Mäurer¹, Elke Müller^{1†}, Thorsten Wolff³, Thomas Rudel^{1†} & Thomas F. Meyer¹

Human host factors required for influenza virus replication

Renate König^{1*}, Silke Stertz^{4*}, Yingyao Zhou⁷, Atsushi Inoue¹, H. -Heinrich Hoffmann⁴, Suchita Bhattacharyya², Judith G. Alamares⁴, Donna M. Tscherne⁴, Mila B. Ortigoza⁴, Yuhong Liang⁴, Qinshan Gao⁴, Shane E. Andrews³, Sourav Bandyopadhyay⁸, Paul De Jesus¹, Buu P. Tu⁷, Lars Pache¹, Crystal Shih¹, Anthony Orth⁷, Ghislain Bonamy⁷, Loren Miraglia⁷, Trey Ideker⁸, Adolfo Garcia-Sastre^{4,5,6}, John A. T. Young², Peter Palese^{4,5}, Megan L. Shaw^{4*} & Sumit K. Chanda^{1*}



nature
2010

Data

results (gene lists) from 4 two-stage RNAi studies

Data

results (gene lists) from 4 two-stage RNAi studies

1. detection

2. confirmation

Data

results (gene lists) from 4 two-stage RNAi studies

1. detection

2. confirmation

study

e.g., one gene

	DL-1	U2OS	A549 _{DE}	A549 _{US}
Detection Screen	1	1	0	0
Confirmation Screen	1	0	0	0
Pattern Code	11	10	00	00

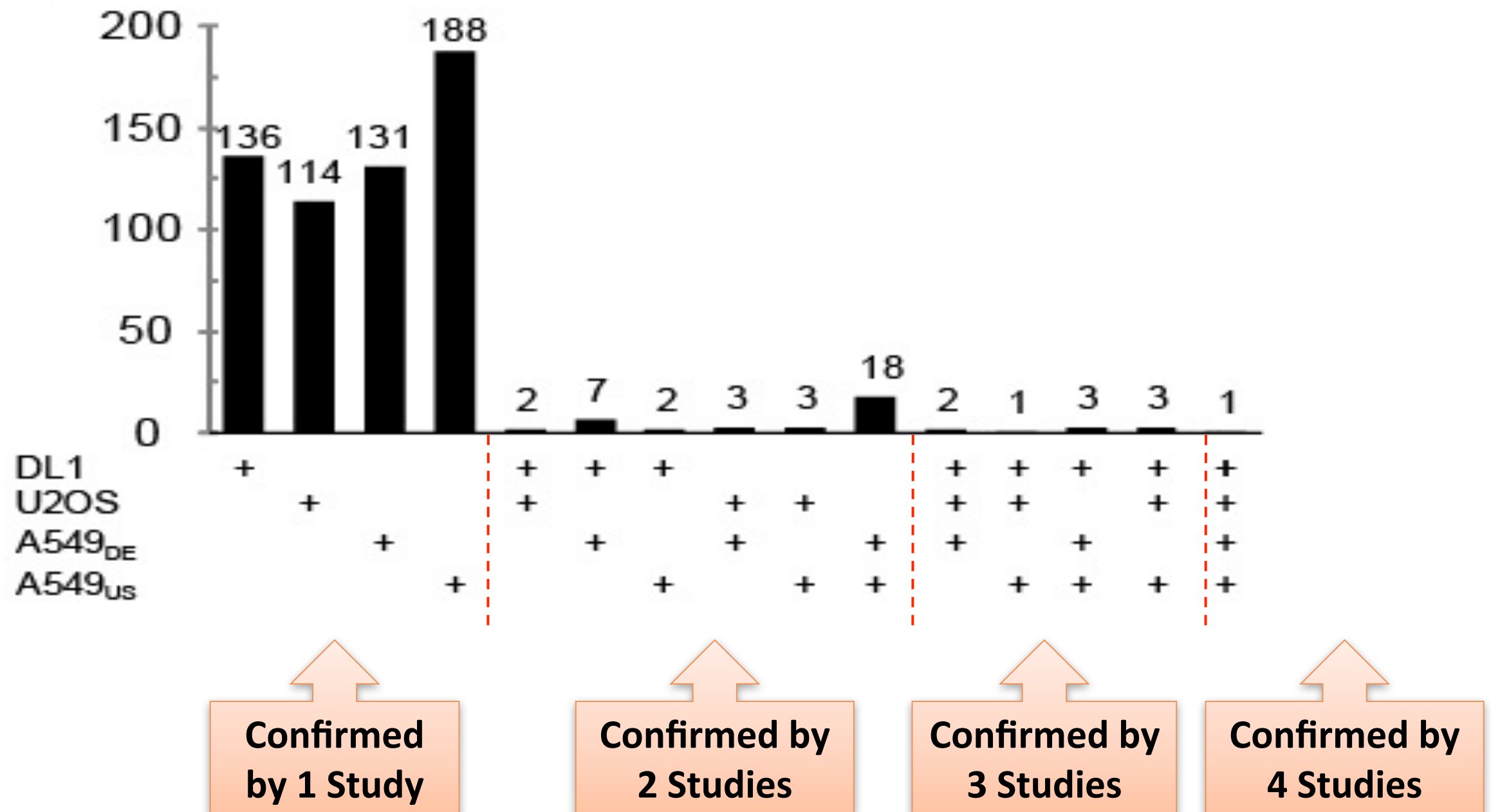
Detection and Confirmation Patterns

Pattern	D=1	U=0	A=0 DE	A=0 US	#grs
1	0	0	0	0	21016
2	10	0	0	0	80
3	11	0	0	0	127
...	...	...	...	...	...
6	11	10	0	0	4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
79	11	11	11	0	3
80	11	11	11	10	0
81	11	11	11	11	1
Total					G=2000

COPG

Agreement among studies is low.

Among the **614** genes confirmed by at least one study:



Is the limited overlap due more to false positive or false negative factors?

6.7% average pairwise overlap of confirmed gene lists is significantly higher than expected by chance

Modeling approach

$$P_{\pi} = \text{Prob}(\text{gene shows detection/confirmation pattern } \pi)$$

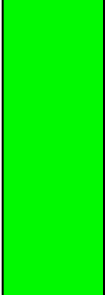
$$n_{\pi} = \#(\text{gene shows detection/confirmation pattern } \pi)$$

Likelihood

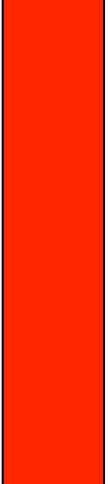
$$L = \prod_{\pi} P_{\pi}^{n_{\pi}}$$

π	DL-1	U2OS	A549 _D E	A549 _U S	n_{π}	P_{π}
1	0	0	0	0	21,016	
2	10	0	0	0	80	
3	11	0	0	0	127	
...	...	...	...	...	...	
6	11	10	0	0	4	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
79	11	11	11	0	3	
80	11	11	11	10	0	
81	11	11	11	11	1	
Total					G = 22000	

Data and latent variables


$$D_{g,s} = 1 \text{ [gene } g \text{ detected in study } s]$$

$$C_{g,s} = 1 \text{ [gene } g \text{ confirmed in study } s]$$

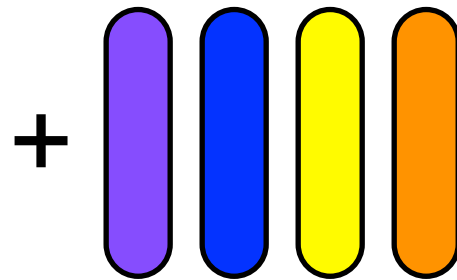

$$I_g = 1 \text{ [} g \text{ involved in flu]}$$

$$A_{g,s} = 1 \text{ [} g \text{ accessible in } s]$$

$$T_{g,s} = \#\{\text{involved, accessible off targets, study } s, \text{ target } g\}$$

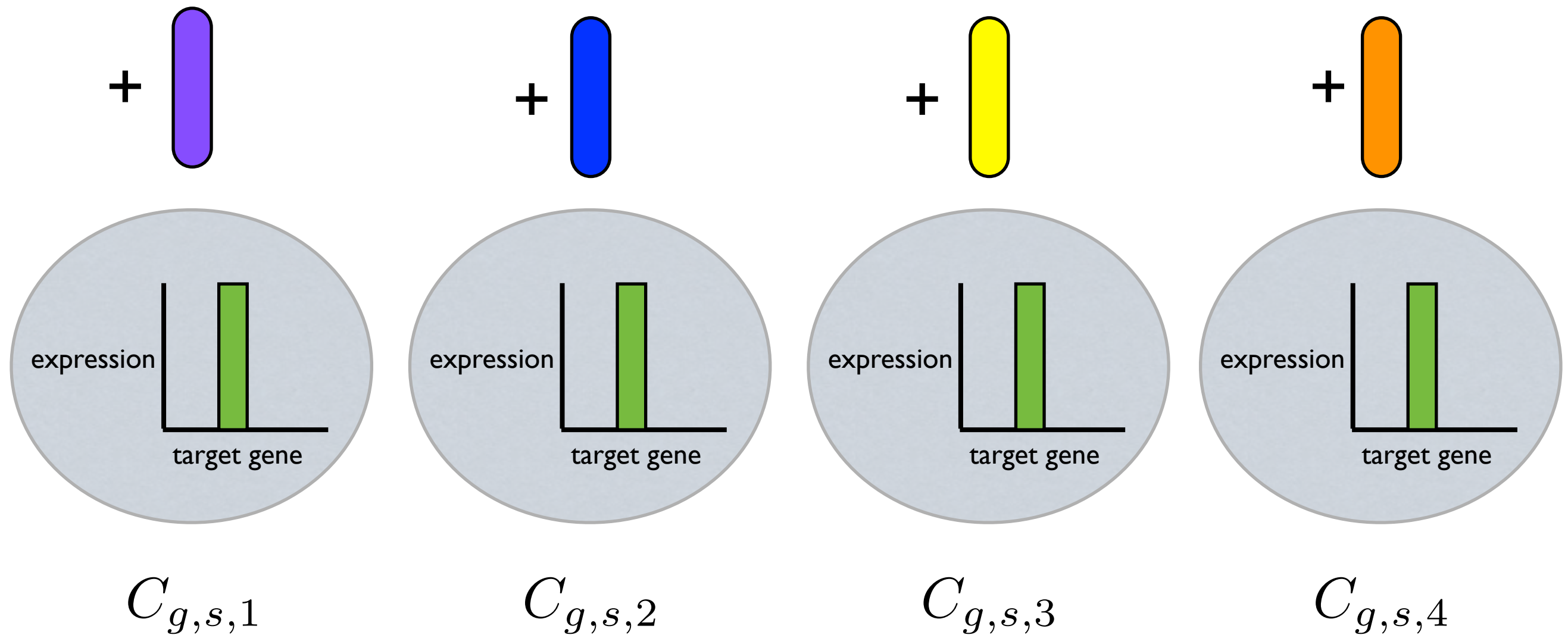
Detection screen

pool of 4 distinct
siRNA's per target gene



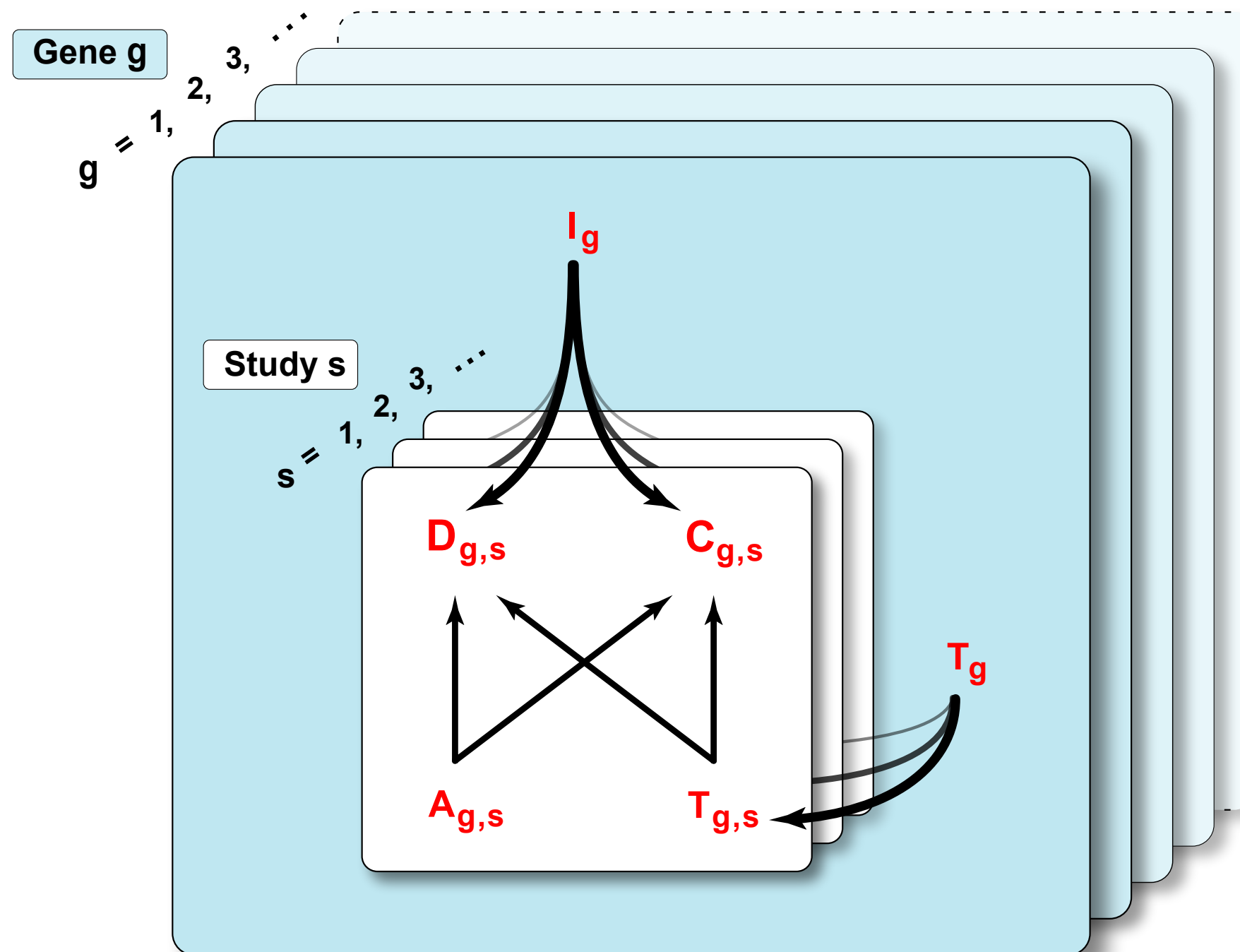
one phenotype call $D_{g,s}$

Confirmation screen



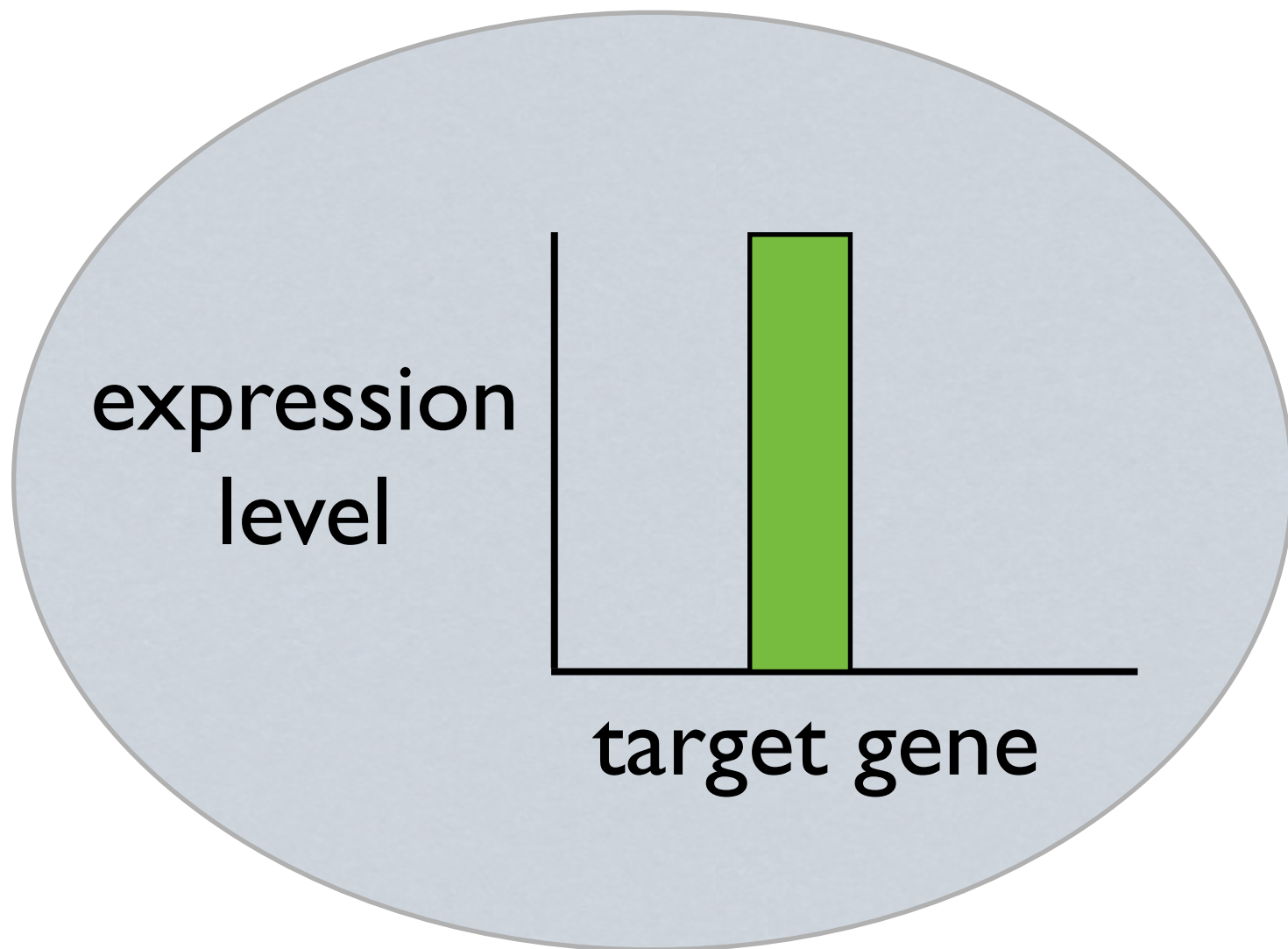
$$C_{g,s} = 1 \left[\sum_k C_{g,s,k} \geq 2 \right]$$

Figure 2. Plate diagram for statistical model



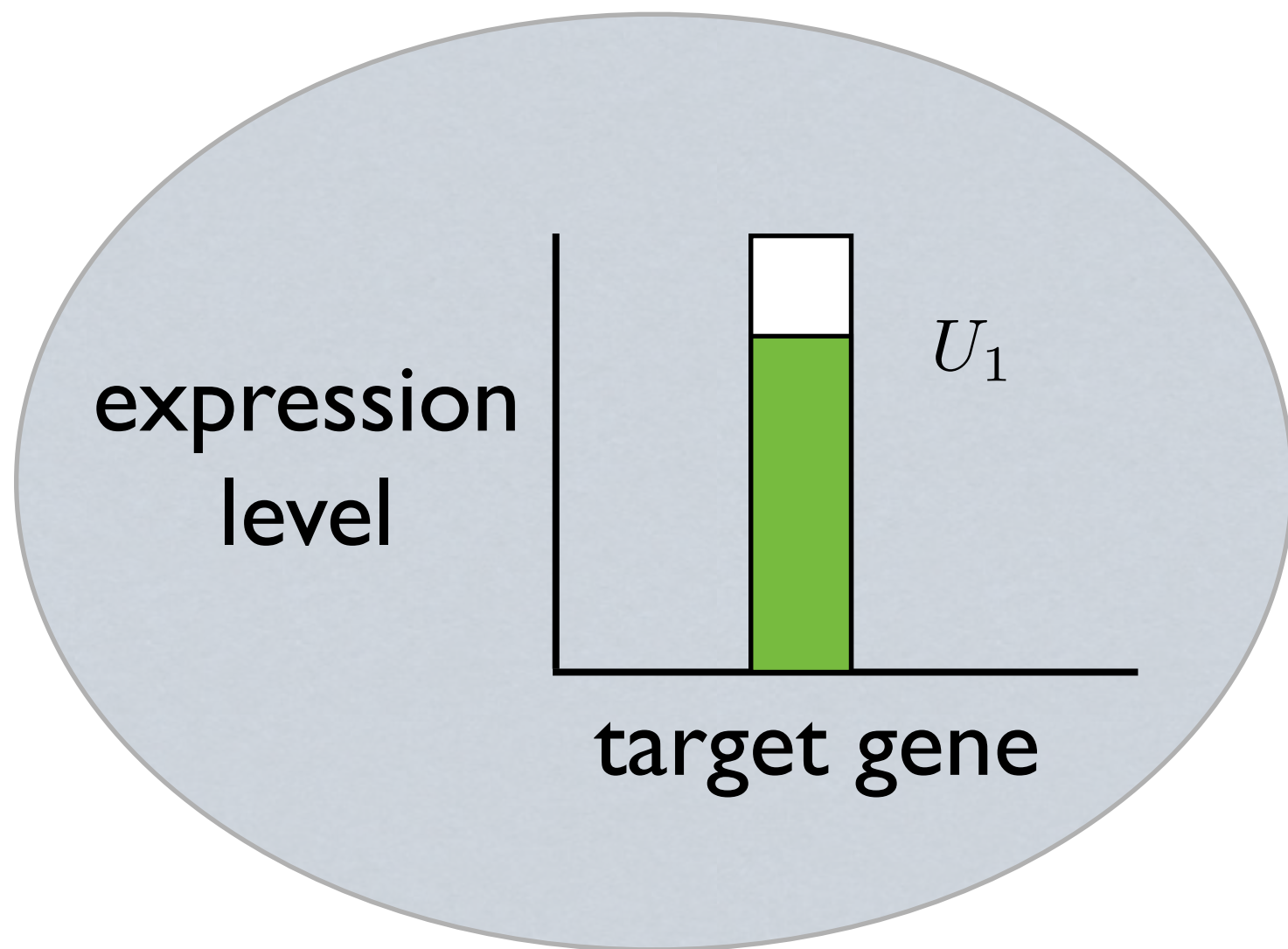
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



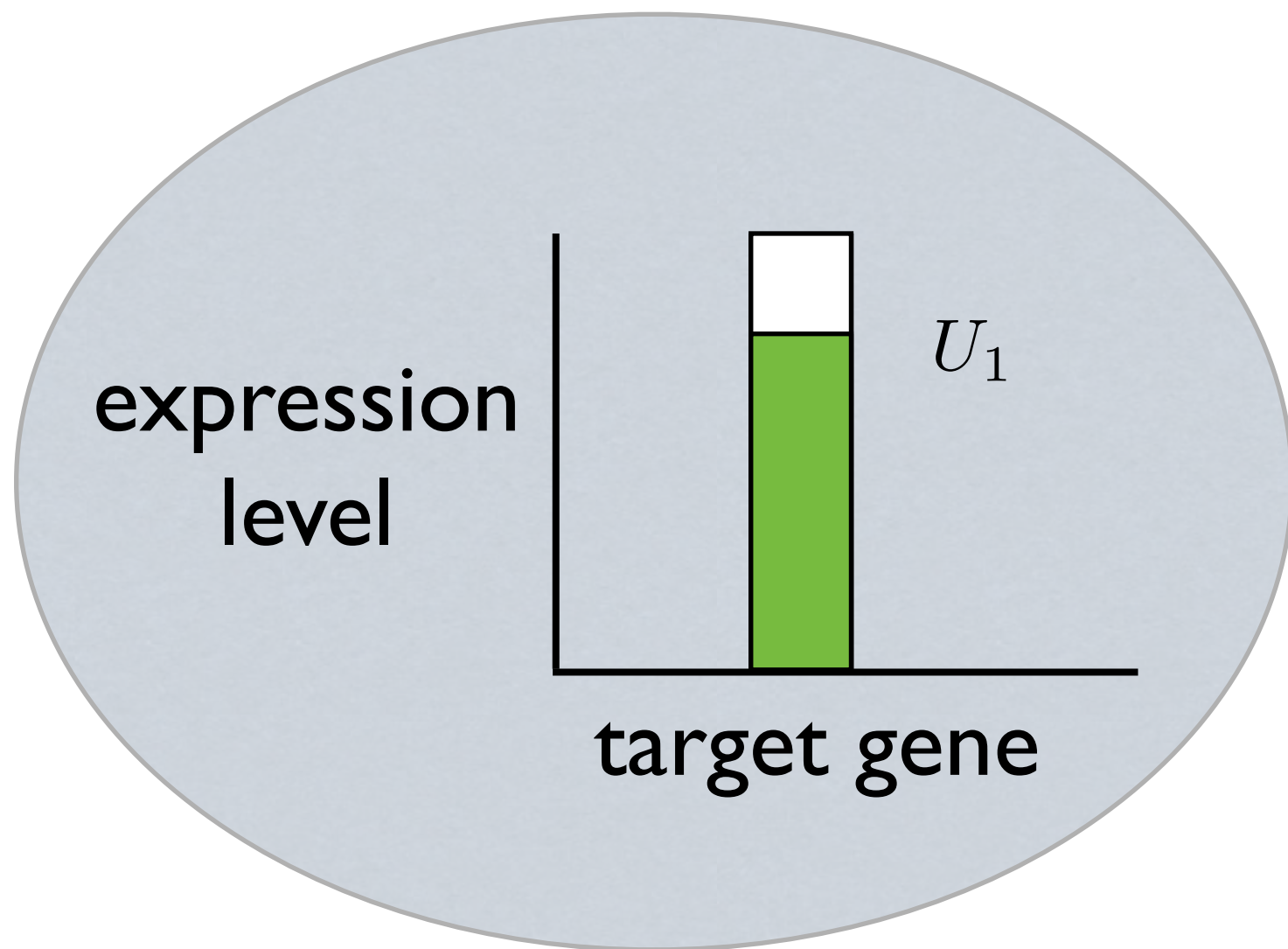
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



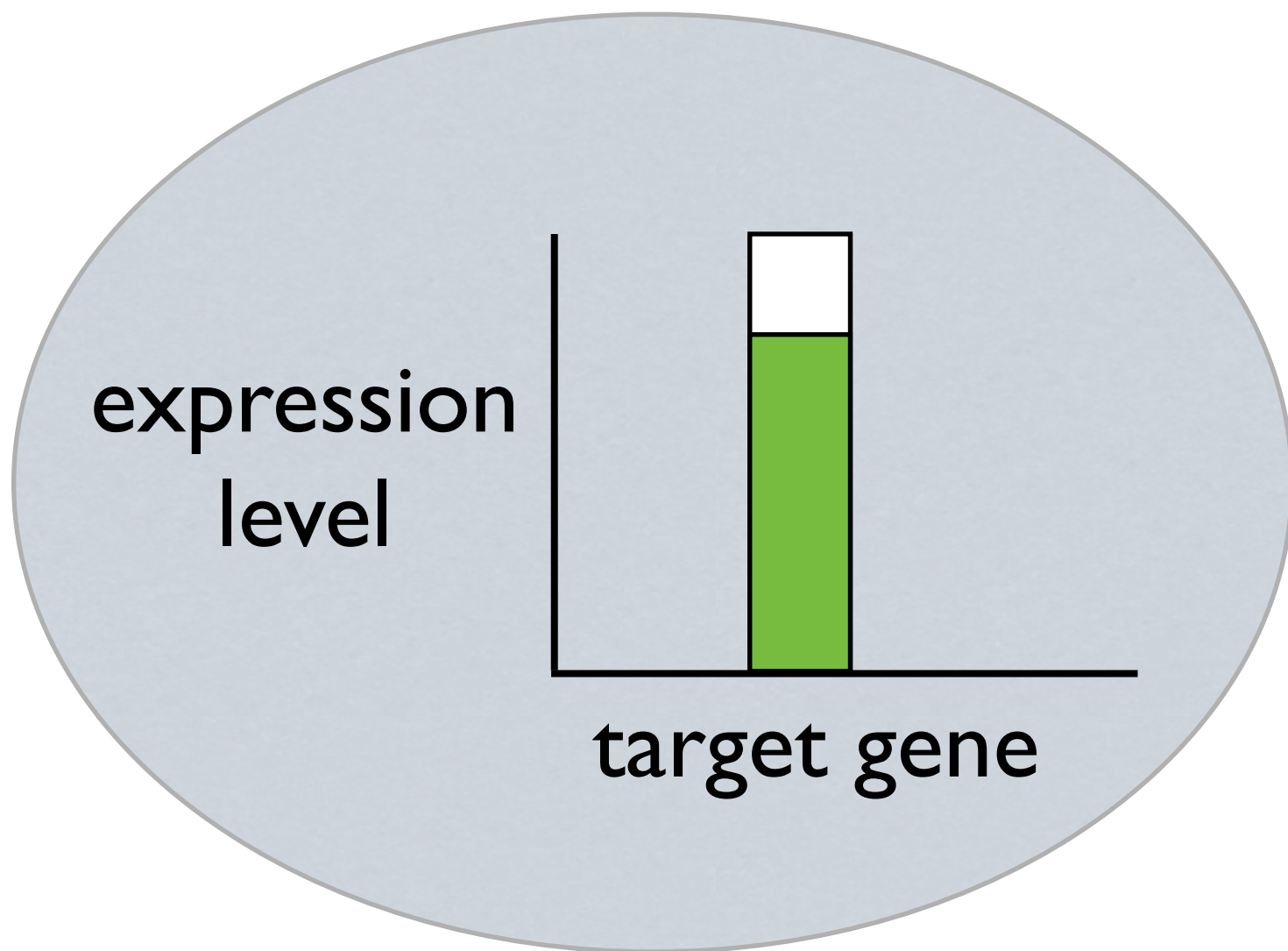
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



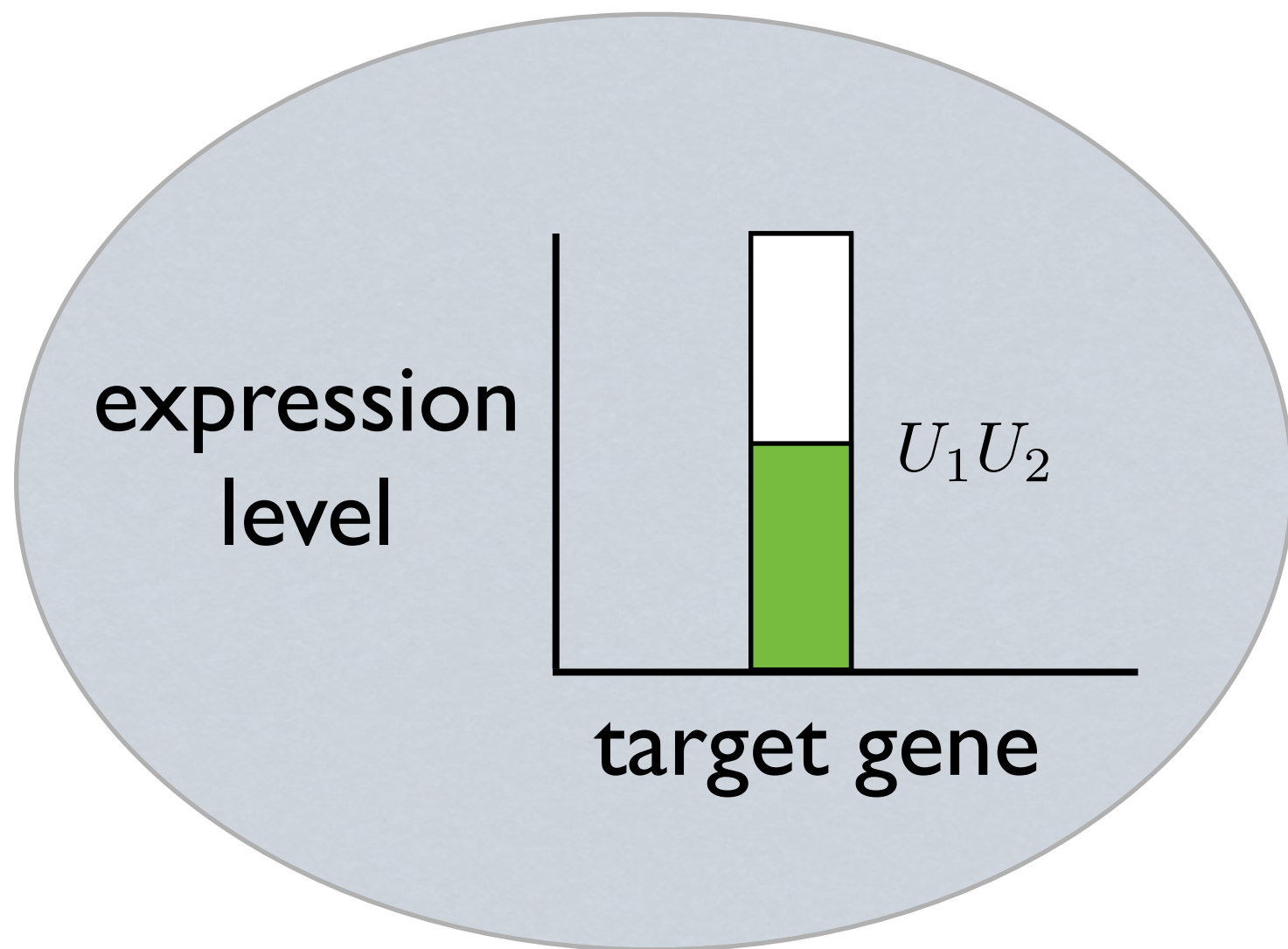
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



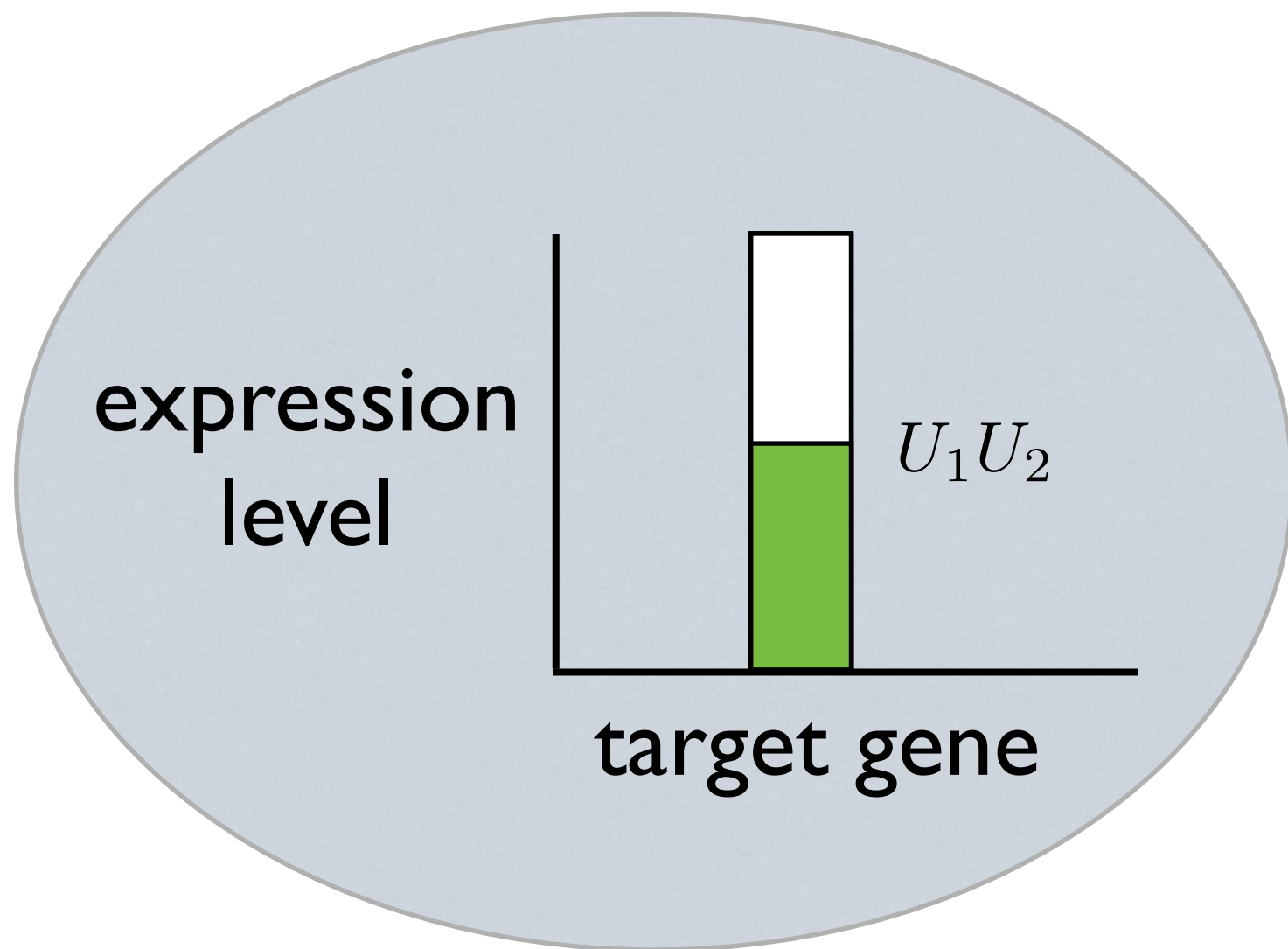
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



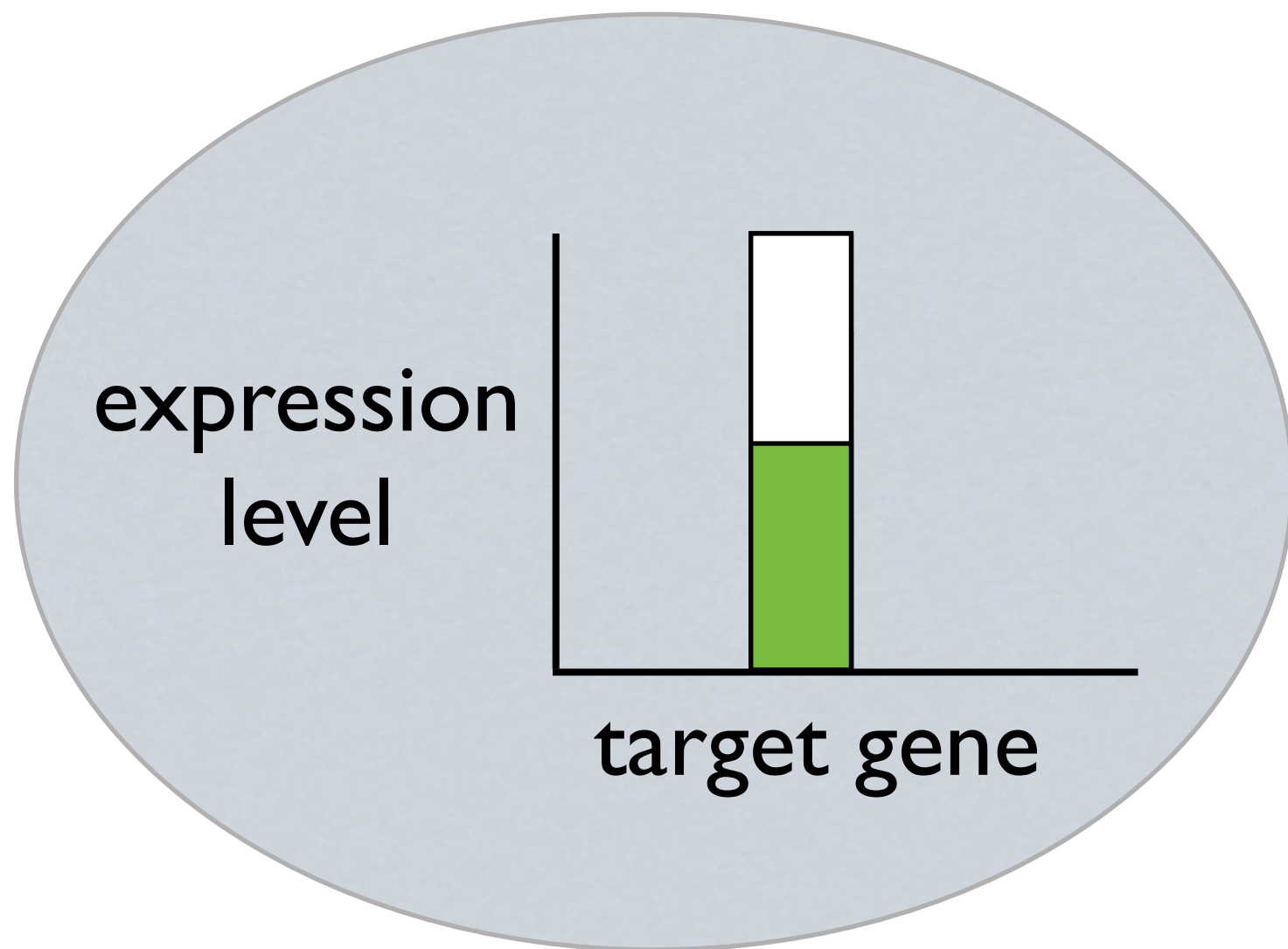
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



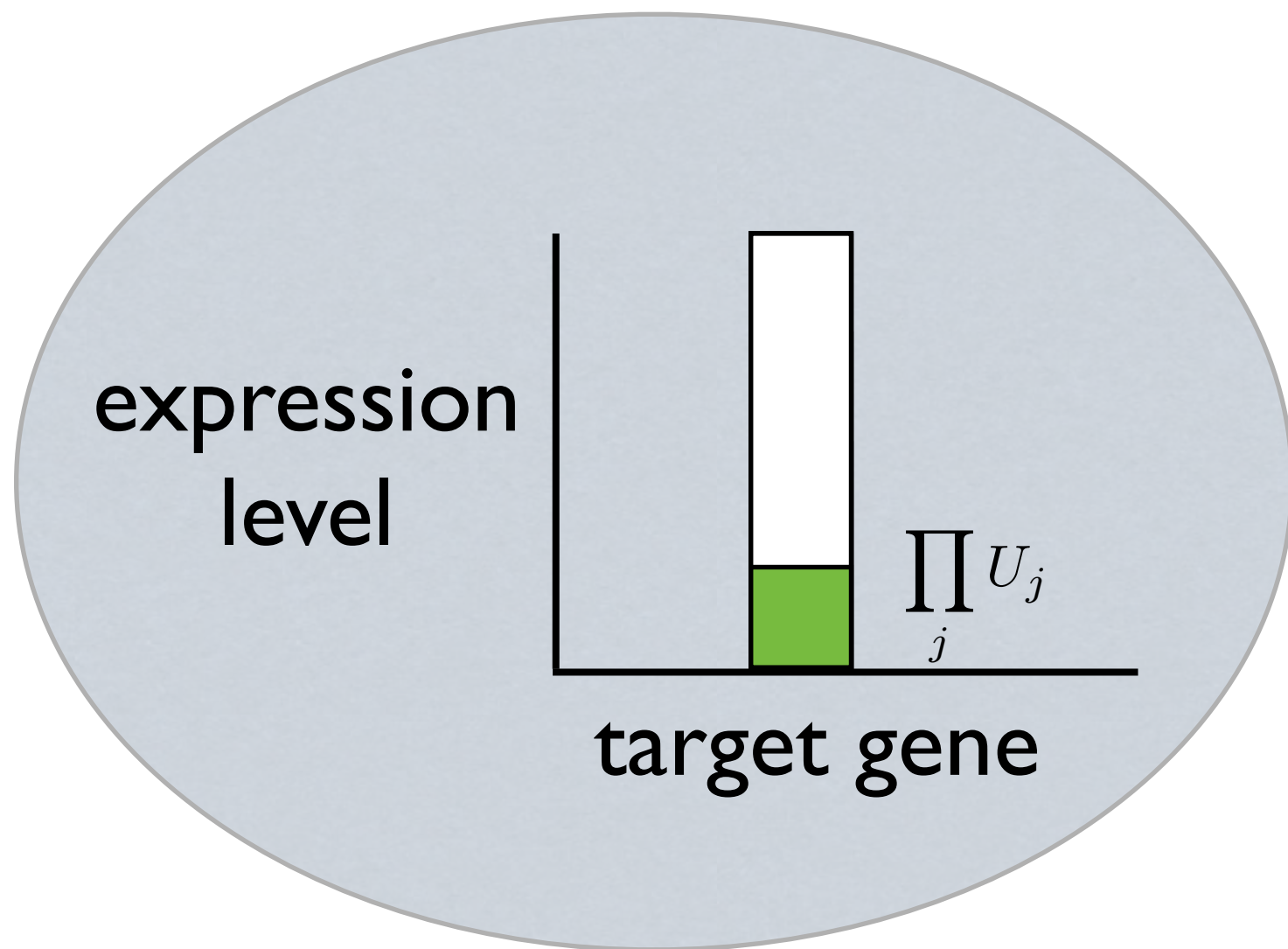
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



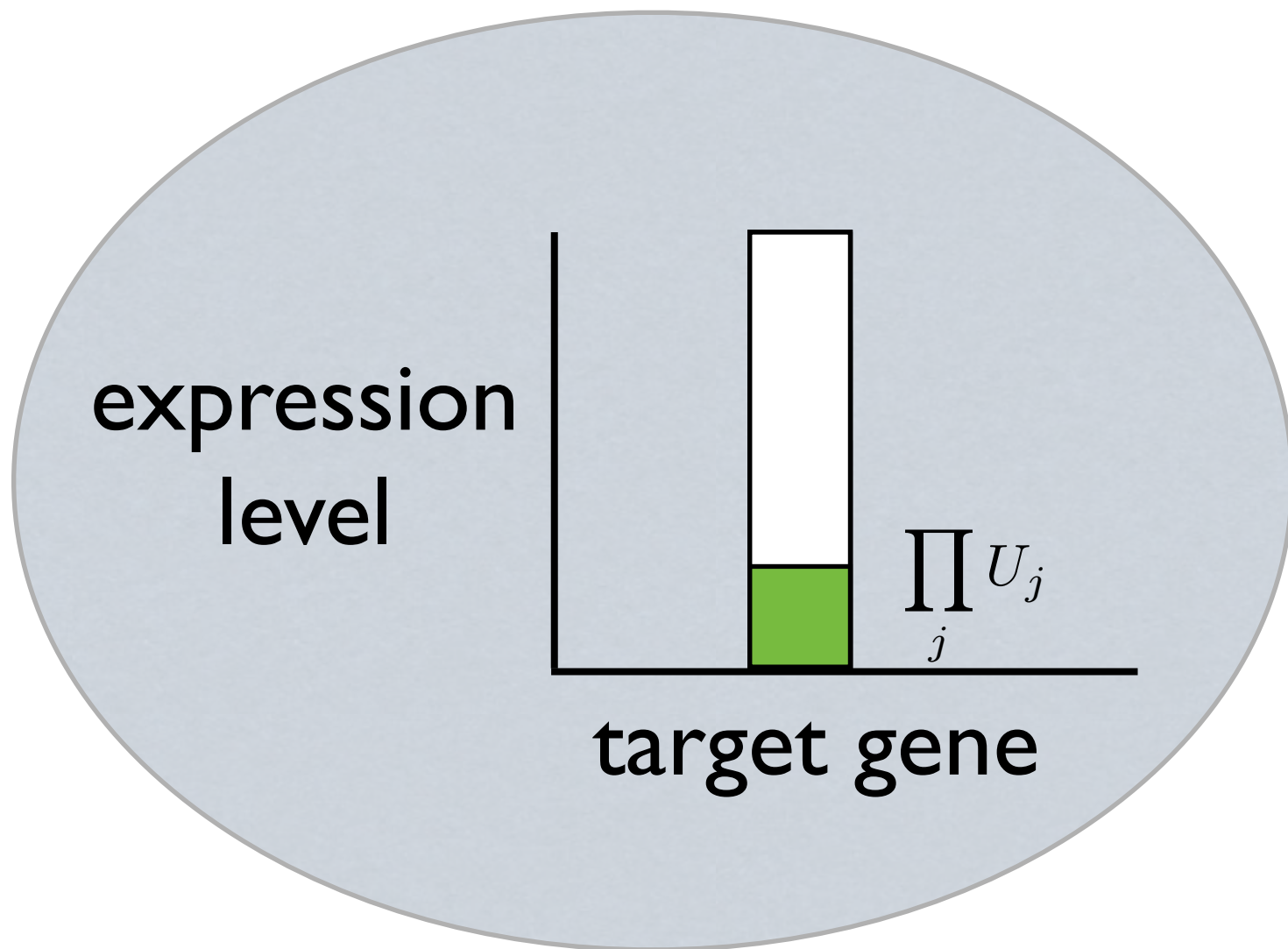
on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



on target knock down model

$$U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1)$$



knock down
effect if

$$\prod_j U_j < \omega$$



threshold
parameter

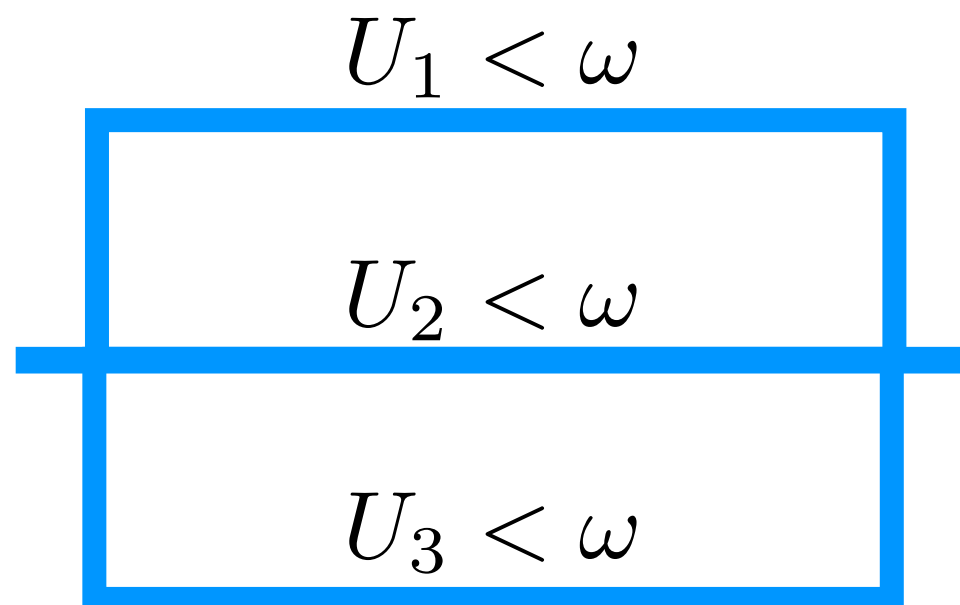
Prob: $1 - G_4 [\log(1/\omega)]$

off target knock down model

$t = \#\{\text{involved, accessible off targets}\}$

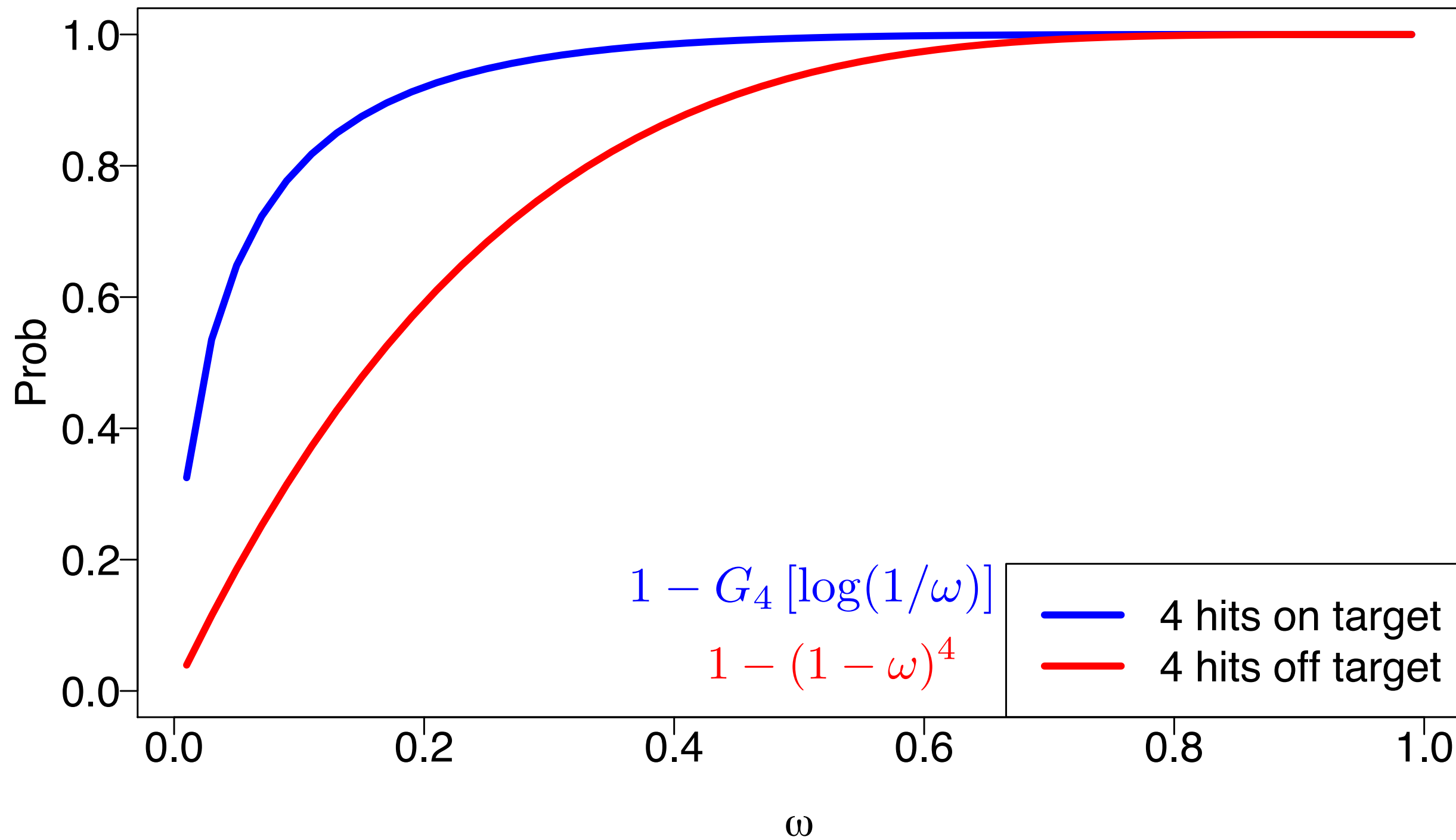
$U_1, U_2, \dots, U_t \sim_{i.i.d.} \text{Uniform}(0, 1)$

e.g. $t = 3$

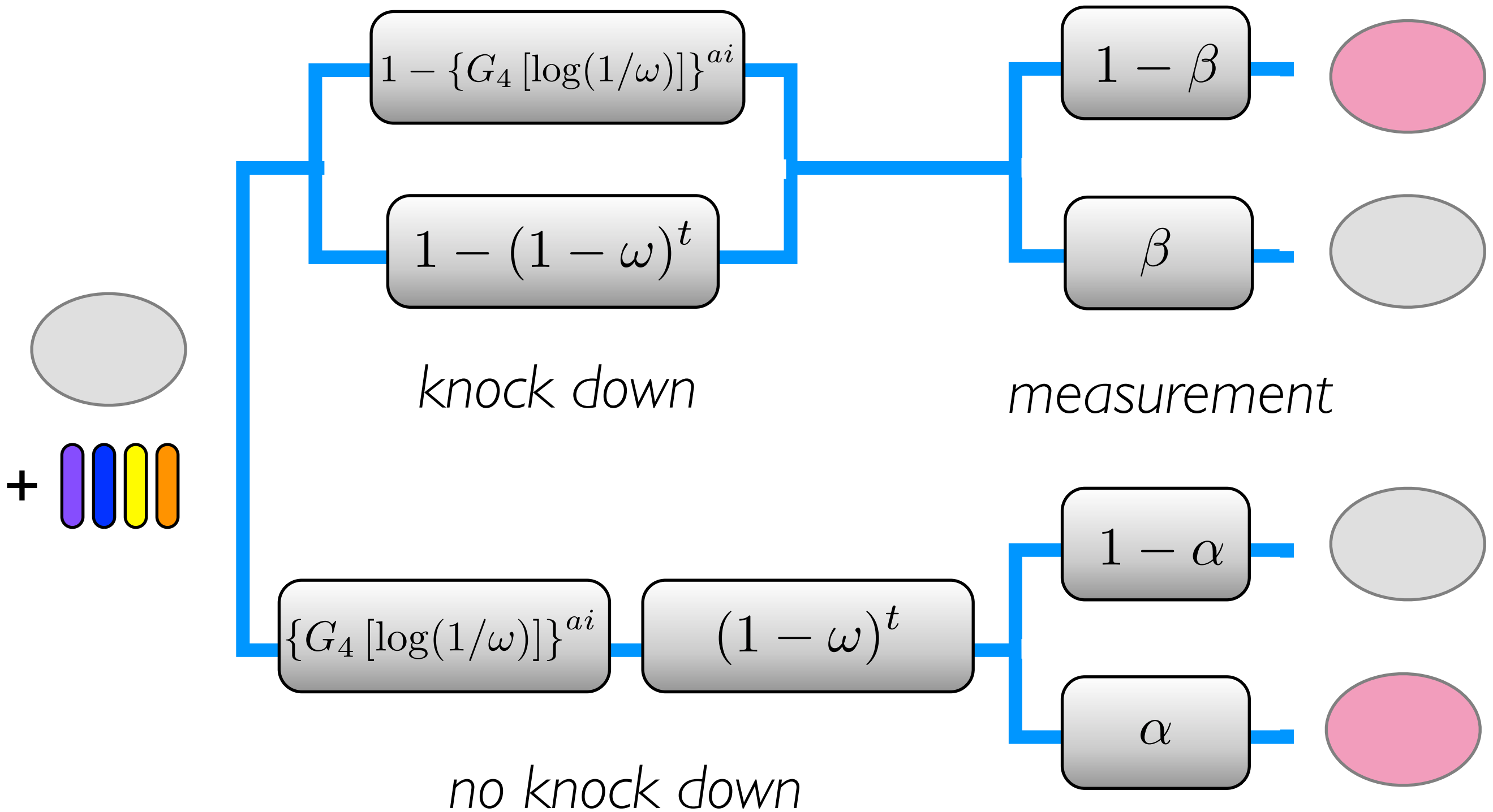


knock down effect
if any $U_j < \omega$

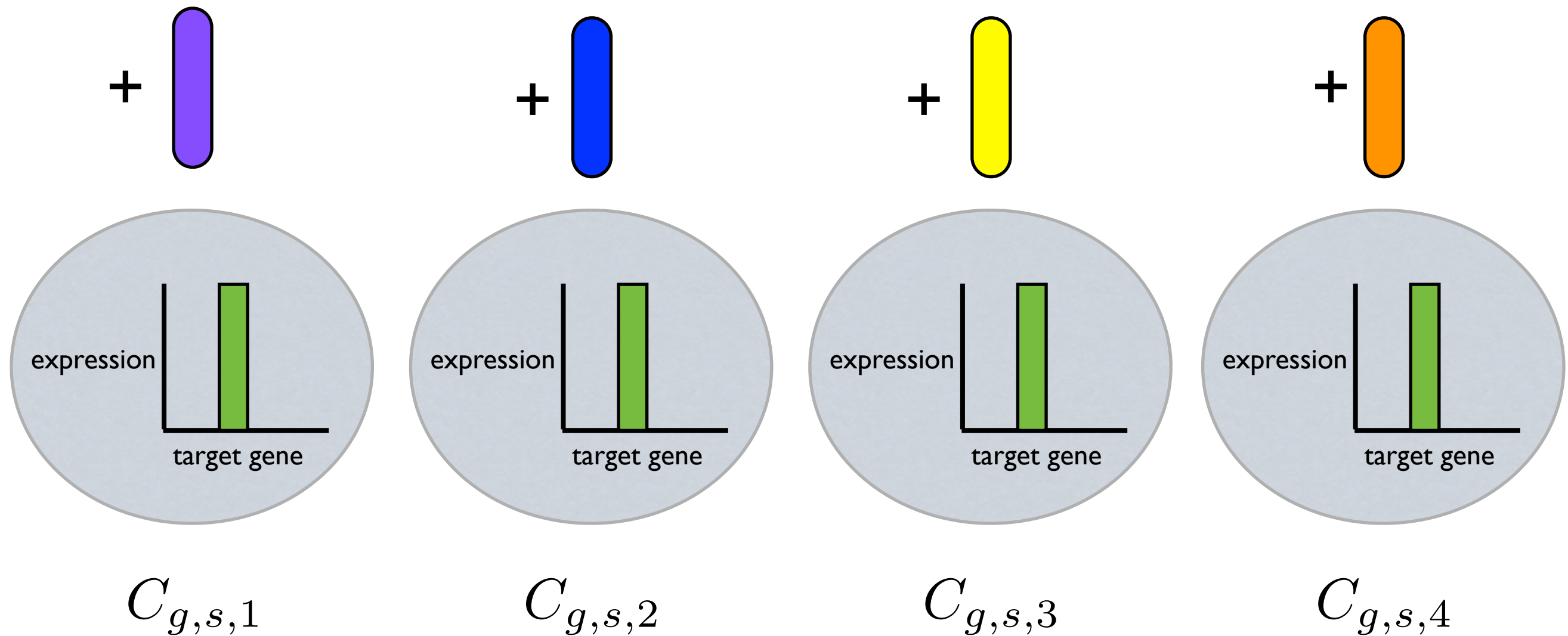
Prob: $1 - (1 - \omega)^t$



detection screen: $D_{g,s} \mid A_{g,s} = a, I_g = i, T_{g,s} = t$

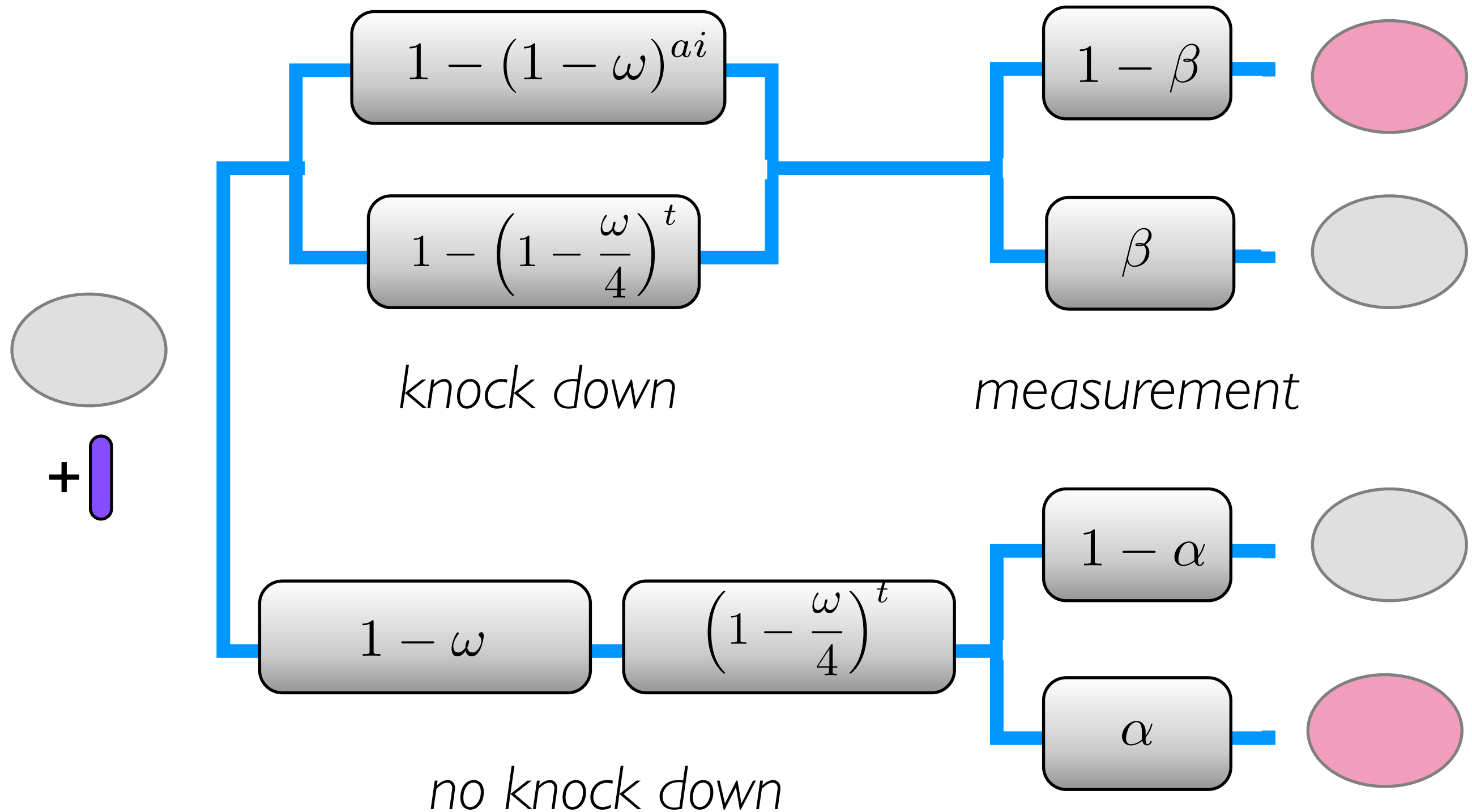


confirmation screen



$$C_{g,s} = 1 \left[\sum_k C_{g,s,k} \geq 2 \right]$$

confirmation screen: $C_{g,s,k} \mid A_{g,s} = a, I_g = i, T_{g,s} = t$



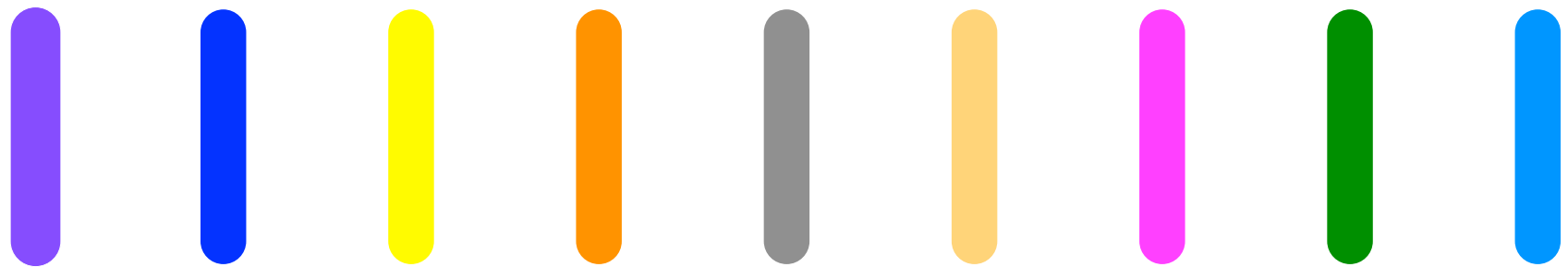
On the number of off targets per siRNA

- very limited data
- libraries overlap among 4 studies

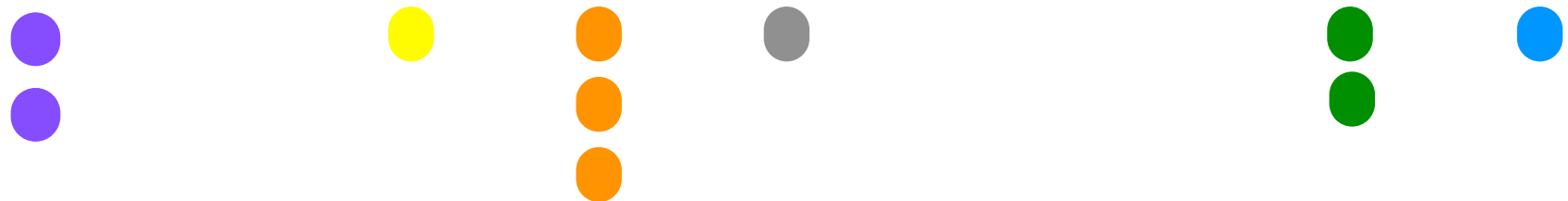
one target gene

K siRNA's available to all studies (on average)

siRNA's



involved
off targets



T_g

one target gene

study s uses 4 siRNA's

siRNA's



involved,
accessible
off targets



$T_{g,s}$

$$T_g = \#\{\text{involved off-targets, target } g \}$$

$$T_g \sim \text{Poisson}(K\theta\nu)$$

$$T_{g,s}|T_g = t \sim \text{Binomial}\left(t, \frac{4\gamma_s}{K}\right)$$

detection

$$T_{g,s,k}|T_{g,s} = t \sim \text{Binomial}\left(t, \frac{1}{4}\right)$$

confirmation

Parameters

- θ = proportion of genome involved in influenza virus replication
- α = false positive measurement error
- β_s = false negative measurement error, study s
- γ_s = rate at which genes are accessible, study s
- ω = knockdown efficiency per siRNA
- ν = average number of off-targets per siRNA .

$$\begin{aligned}
I_g &\sim \text{Bernoulli}(\theta) \\
A_{g,s} &\sim \text{Bernoulli}(\gamma_s) \\
T_g &\sim \text{Poisson}(K\theta\nu) \\
T_{g,s} \mid [T_g = t] &\sim \text{Binomial}(t, 4\gamma_s/K) \\
D_{g,s} \mid [I_g = i, A_{g,s} = a, T_{g,s} = t] &\sim \text{Bernoulli} \left[1 - \beta_s + (\alpha + \beta_s - 1) [G_{4,1}(-\log \omega)]^{ai} (1 - \omega)^t \right] \\
C_{g,s,k} \mid [I_g = i, A_{g,s} = a, T_{g,s} = t] &\sim \text{Bernoulli} \left[1 - \beta_s + (\alpha + \beta_s - 1)(1 - \omega)^{ai} (1 - \omega/4)^t \right]
\end{aligned}$$

Calculating pattern probabilities:

$$P_{\pi} = \sum_{i,a} P_{\pi}(i, a)$$

$$P_{\pi}(i, a) = P(\pi | I_g = i, \{A_{g,s}\}_{s=1}^4 = a) .$$

$$\begin{aligned} P_{\pi}(i, a) &= \sum_{t=0}^{\infty} P(T_g = t) P(\pi | I_g = i, \{A_{g,s}\}_{s=1}^4 = a, T_g = t) \\ &= \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^4 P(\pi_s | I_g = i, A_{g,s} = a_s, T_g = t) \quad \pi = \bigcap_s \pi_s. \\ &= \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^4 \sum_{u=0}^t B_s(t, u) P(\pi_s | I_g = i, A_{g,s} = a_s, T_{g,s} = u) \\ &= \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^4 \sum_{u=0}^t B_s(t, u) Q_{s,i,a_s,u} \end{aligned}$$

$$Q_{s,i,a_s,u} = P(\pi_s | I_g = i, A_{g,s} = a_s, T_{g,s} = u)$$

Lemma:

$$Q_{s,i,a_s,u} = \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} (\xi_1^p \xi_2^q)^u$$

$$\xi_1 = 1 - \omega$$

$$\xi_2 = 1 - \frac{\omega}{4}$$

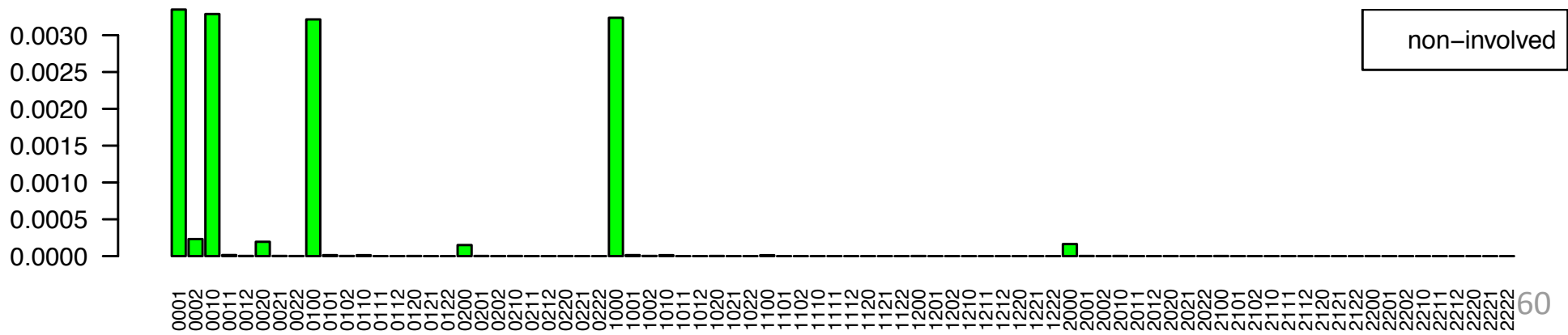
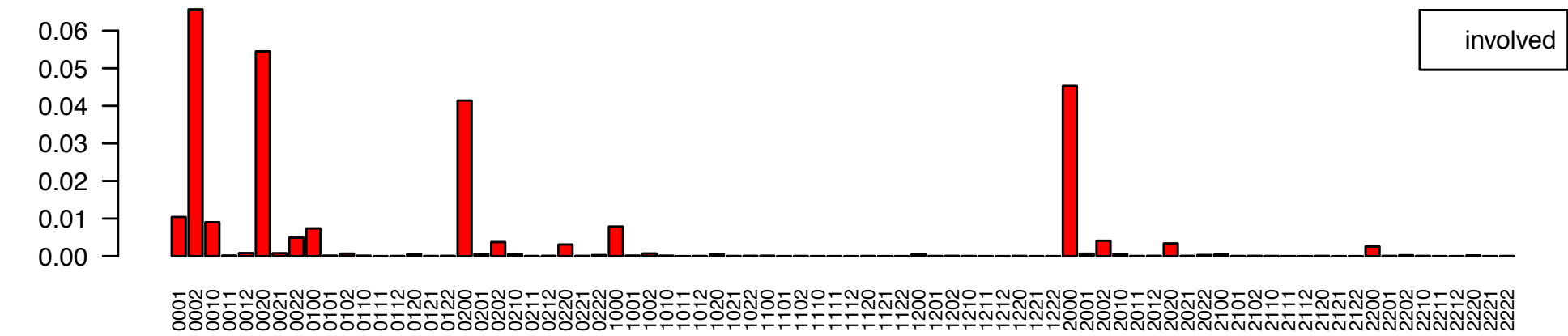
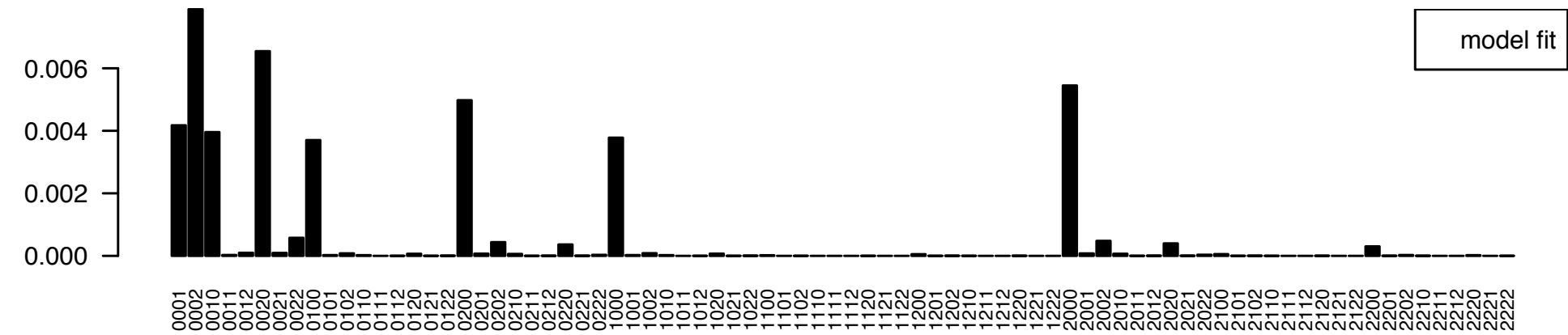
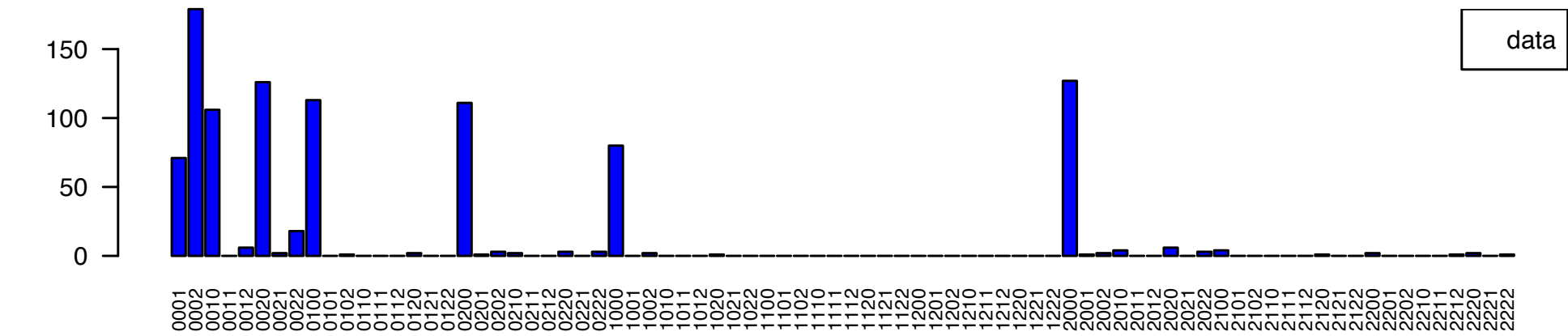
$$\begin{aligned}
\sum_{u=0}^t B_s(t, u) Q_{s,i,a_s,u} &= \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} \sum_{u=0}^t (\xi_1^p \xi_2^q)^u B_s(t, u) \\
&= \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} \left(1 - \frac{4\gamma_s}{K} + \frac{4\gamma_s}{K} (\xi_1^p \xi_2^q) \right)^t \\
&= \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} e_{s,p,q}^t,
\end{aligned}$$

$$e_{s,p,q} = 1 - \frac{4\gamma_s}{K} + \frac{4\gamma_s}{K} (\xi_1^p \xi_2^q)$$

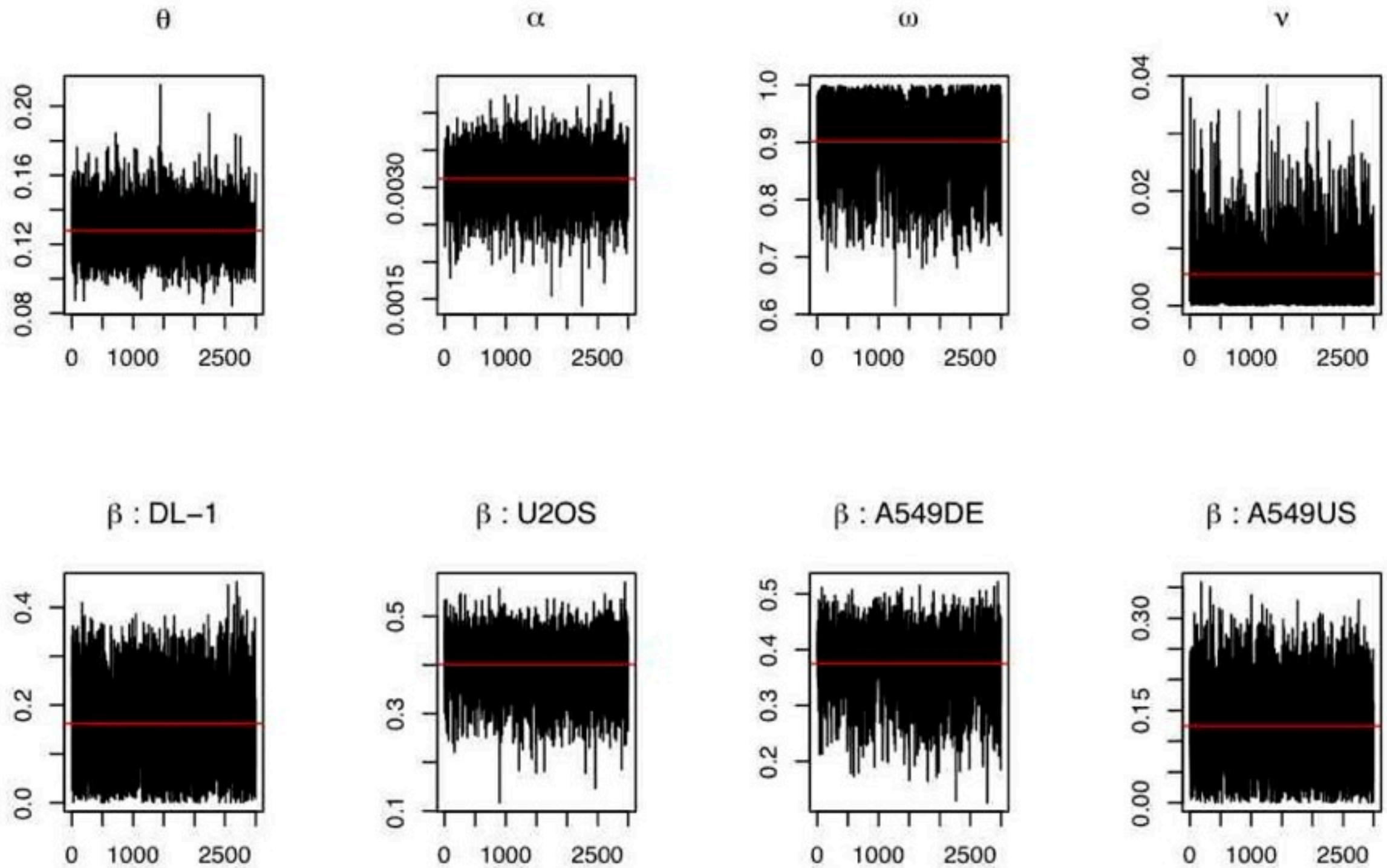
$$\begin{aligned}
P_\pi(i, a) &= \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^4 \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} e_{s,p,q}^t \\
&= \sum_{t=0}^{\infty} \text{Po}(t) \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \left(\prod_{s=1}^4 e_{s,p_s,q_s} \right)^t \\
&= \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \sum_{t=0}^{\infty} \text{Po}(t) \left(\prod_{s=1}^4 e_{s,p_s,q_s} \right)^t \\
&= \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \exp \left\{ -K\theta\nu \left(1 - \prod_s e_{s,p_s,q_s} \right) \right\}
\end{aligned}$$

Model fitting

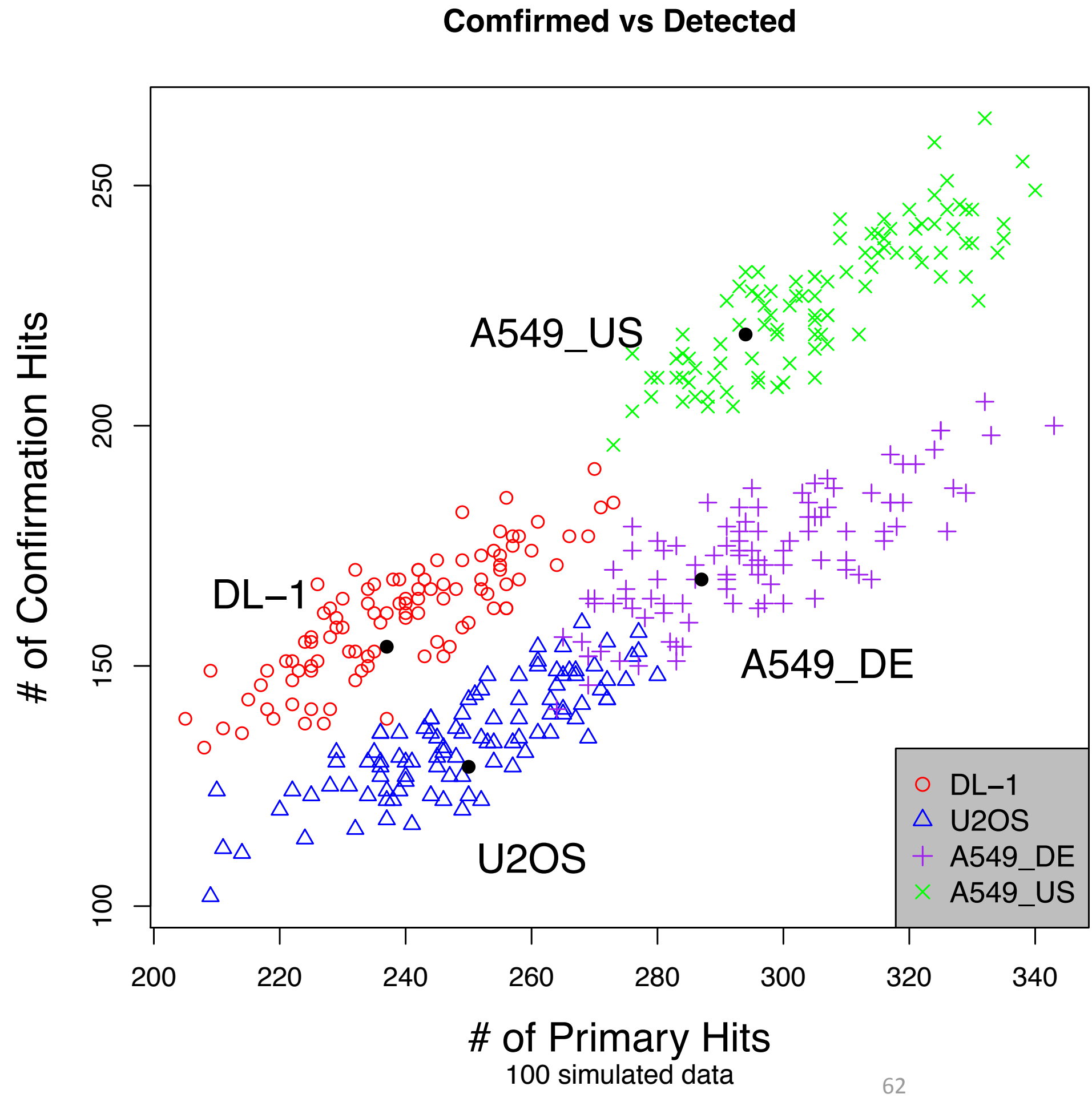
- 12 parameters
- numerical (*nlminb* in R) (point estimation)
- extensive code testing
- MCMC (Bayes under flat prior)
(induced parameters and prediction)



MCMC looks good



Posterior
predictive
checks look
good



Cross validation

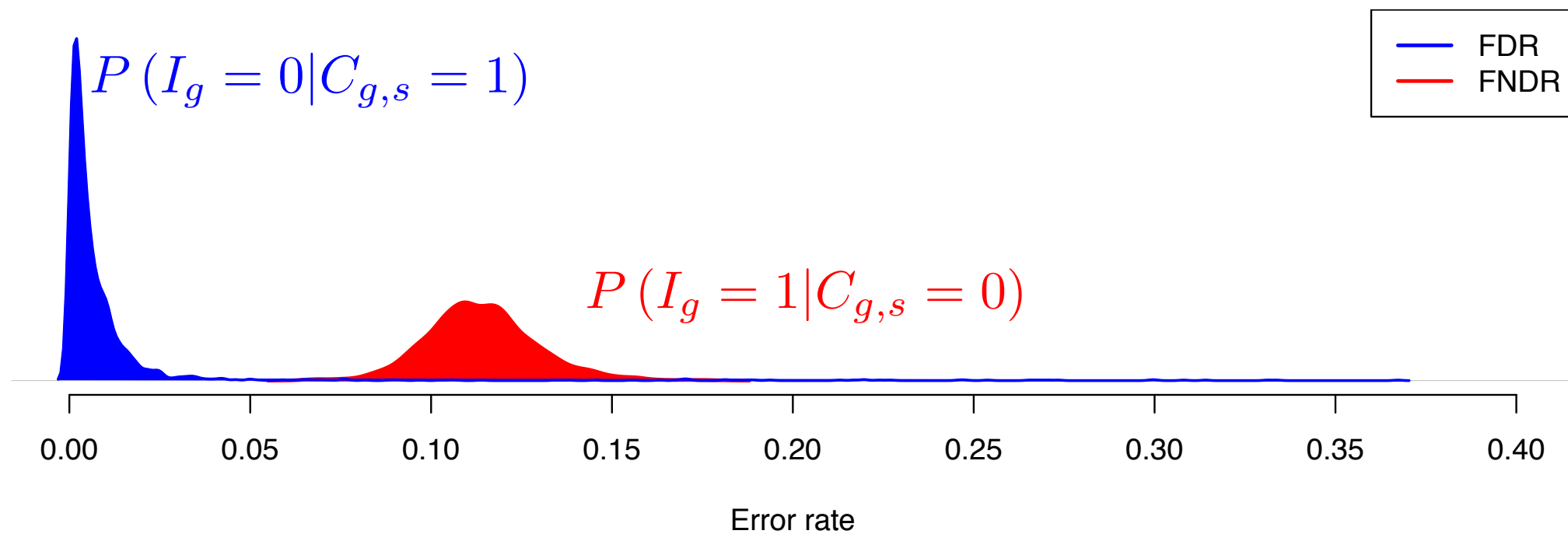
Table S3-5: Predicted number of extra genes confirmed by a 4th study based on modeling the other three studies.

Leave Out	Predicted Additional	95% Prediction Interval	Observed Additional
DL-1	133	(76, 207)	136
U2OS	128	(75, 199)	114
A549US	144	(89, 212)	188
A549DE	156	(80, 284)	131

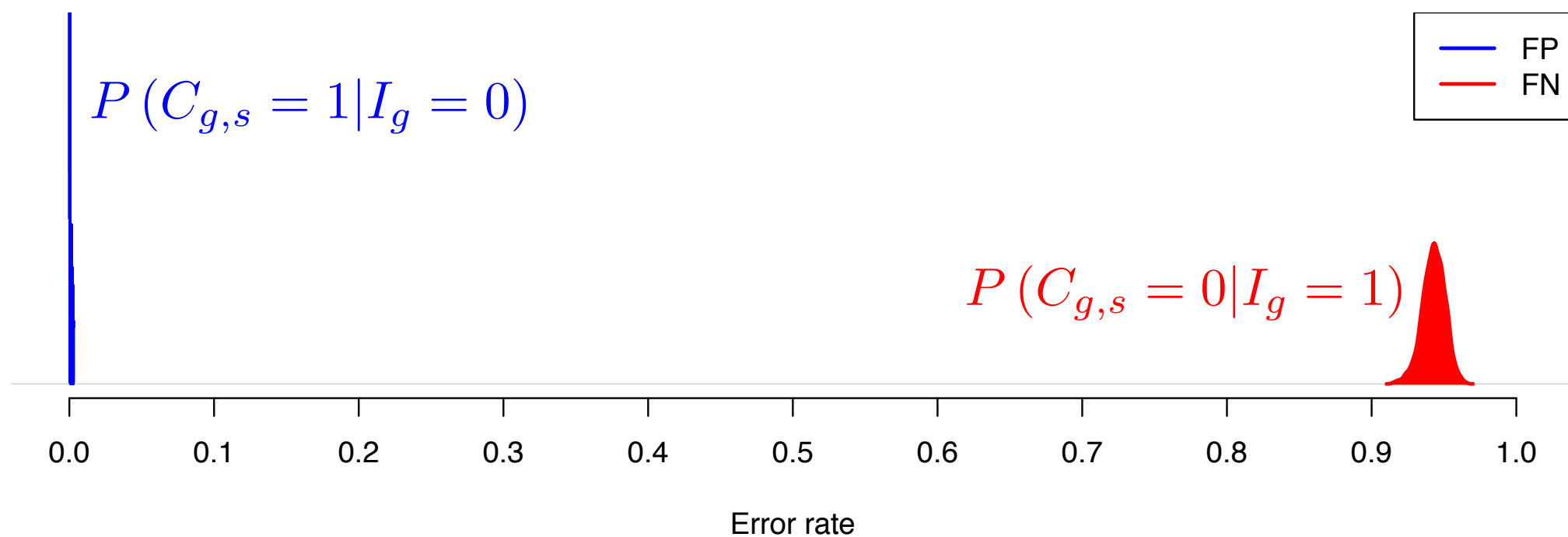
Point estimates

	MLE	MEAN
θ	0.128	0.128
α	0.003	0.003
β_1	0.083	0.164
β_2	0.340	0.400
β_3	0.312	0.375
β_4	0.038	0.125
γ_1	0.063	0.072
γ_2	0.094	0.107
γ_3	0.113	0.127
γ_4	0.084	0.095
ω	0.809	0.902
ν	0.000	0.006

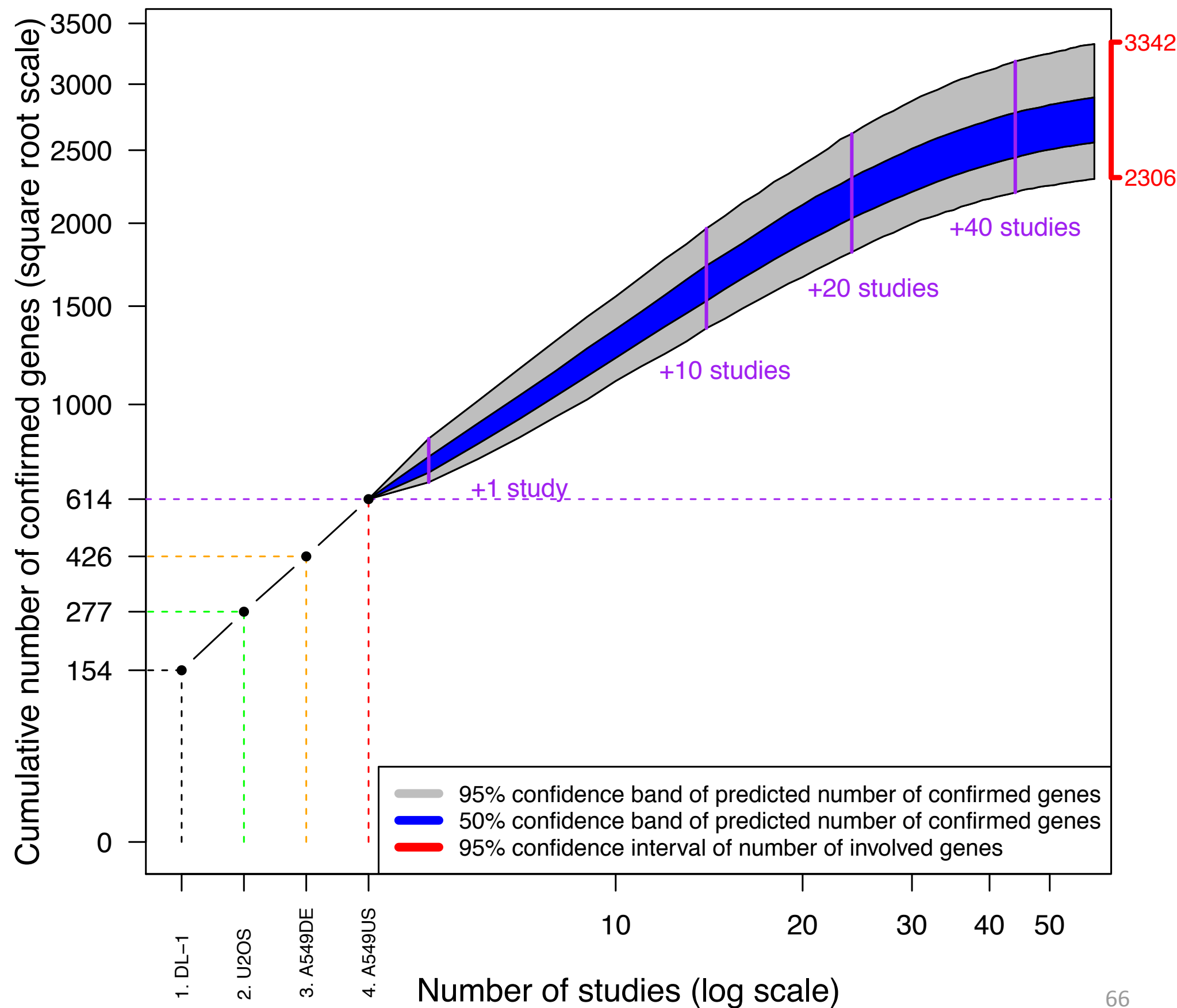
Density of False Discovery Rate (FDR) and False Non-Discovery Rate (FNDR)



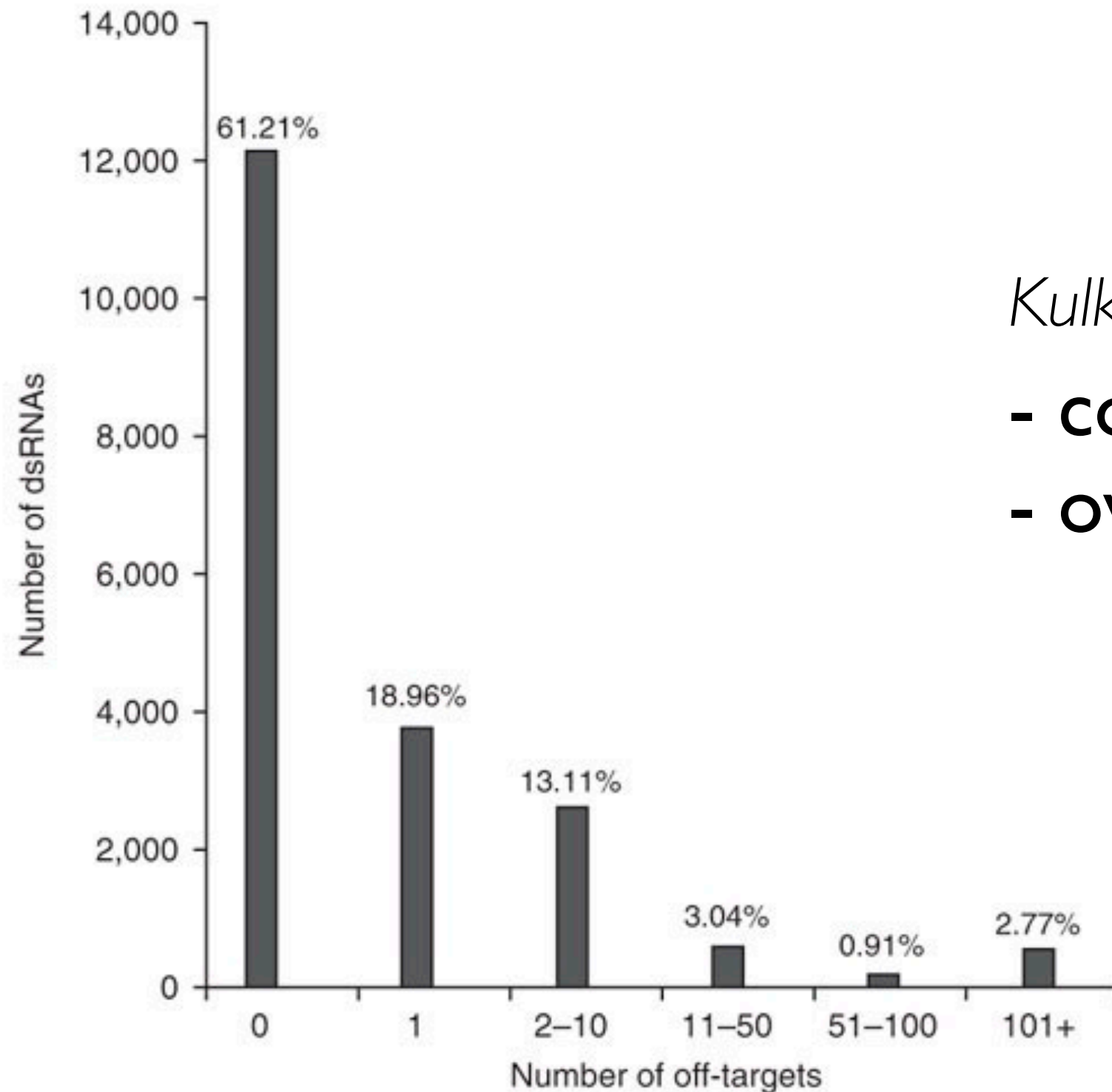
Density of False Positive Rate (FP) and False Negative Rate (FN)



Predicted number of confirmed genes



What about low estimated off target rate ??



Kulkarni et al. 2006, Nat. Meth.

- computational predictions
- overestimate

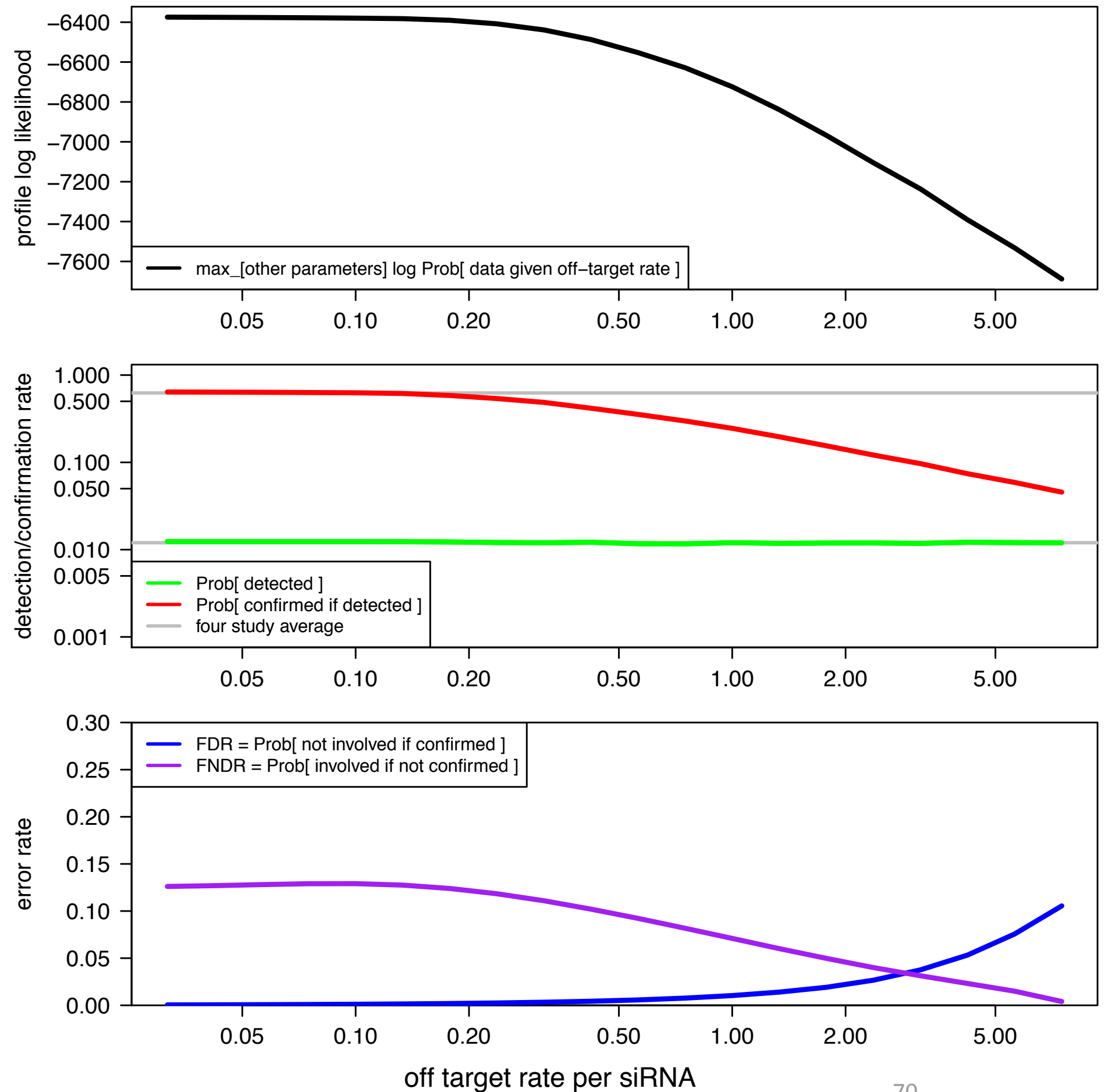
very hard to extend to Negative Binomial

$$\begin{aligned}
 P_{\pi}(i, a) &= \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^4 \sum_{p=0}^1 \sum_{q=0}^4 b_{s,p,q} e_{s,p,q}^t \\
 &= \sum_{t=0}^{\infty} \text{Po}(t) \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \left(\prod_{s=1}^4 e_{s,p_s,q_s} \right)^t \\
 &= \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \sum_{t=0}^{\infty} \text{Po}(t) \left(\prod_{s=1}^4 e_{s,p_s,q_s} \right)^t \\
 &= \sum_{p_1=0}^1 \sum_{p_2=0}^1 \sum_{p_3=0}^1 \sum_{p_4=0}^1 \sum_{q_1=0}^4 \sum_{q_2=0}^4 \sum_{q_3=0}^4 \sum_{q_4=0}^4 \left(\prod_{s=1}^4 b_{s,p_s,q_s} \right) \exp \left\{ -K\theta\nu \left(1 - \prod_s e_{s,p_s,q_s} \right) \right\}
 \end{aligned}$$

Sensitivity analysis

- large $K \iff$ independent studies
- simpler likelihood
- separate implementation with Negative Binomial gives essentially the same fits

Profile analysis



Summary

1. gene-level agreement among studies
2. functional category analysis
3. protein interaction analysis

