Massively parallel assays

Justin B. Kinney
Professor and Chair
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

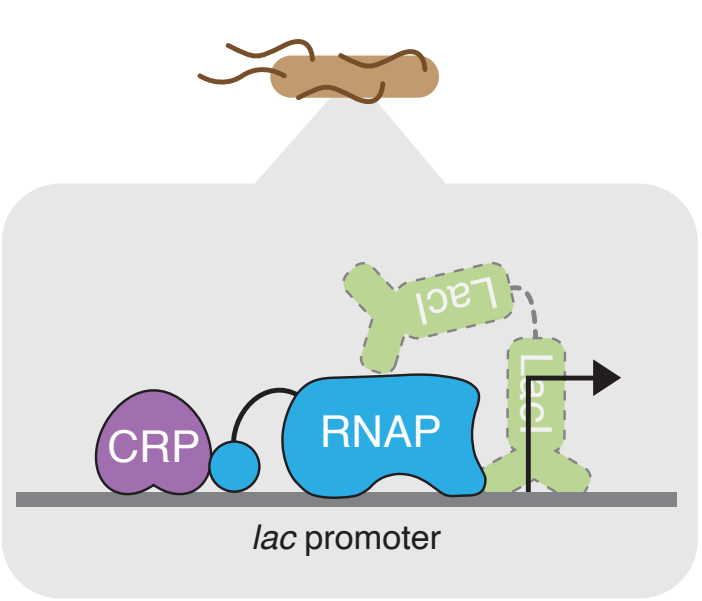
Advanced Sequencing Technologies & Bioinformatics Analysis
CSHL Course
19 November 2025

My lab studies the biophysical mechanisms of gene regulation by quantitatively measuring and modeling sequence-function relationships.

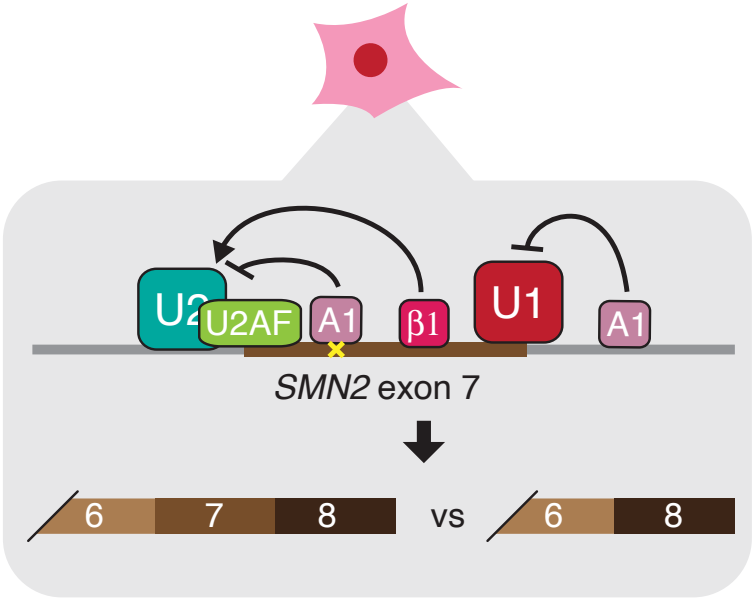
Our work involves a tightly knit combination of experiment, computation, and mathematical theory.

Our **experimental work** uses massively parallel reporter assays (MPRAs) to measure the effects that large numbers of different mutations in gene regulatory sequences have on gene expression. We pursue this experimental work in two biological contexts:

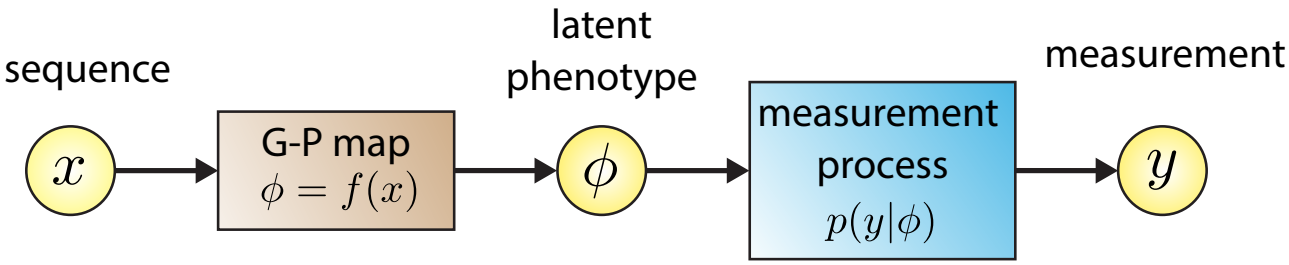
transcriptional regulation in bacteria



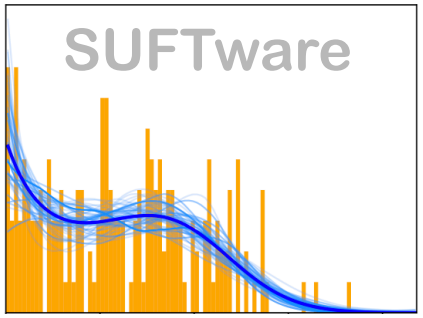
alternative mRNA splicing in human cells



Our **theoretical and computational work** develops methods for analyzing the data produced by MPRAs and other highly multiplexed assays. We aim to extract biophysically meaningful models of regulatory sequence function, but also to understand the quantitative nature of sequence-function relationships more broadly.



These efforts include devising new mathematical and computational approaches, as well as developing robust and efficient software for use by the larger genomics community.



https://suftware.readthedocs.io/

https://logomaker.readthedocs.io/

http://mavenn.readthedocs.io/

Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence

---

Kinney JB, Murugan A, Callan CG, Cox EC
PNAS (2010)

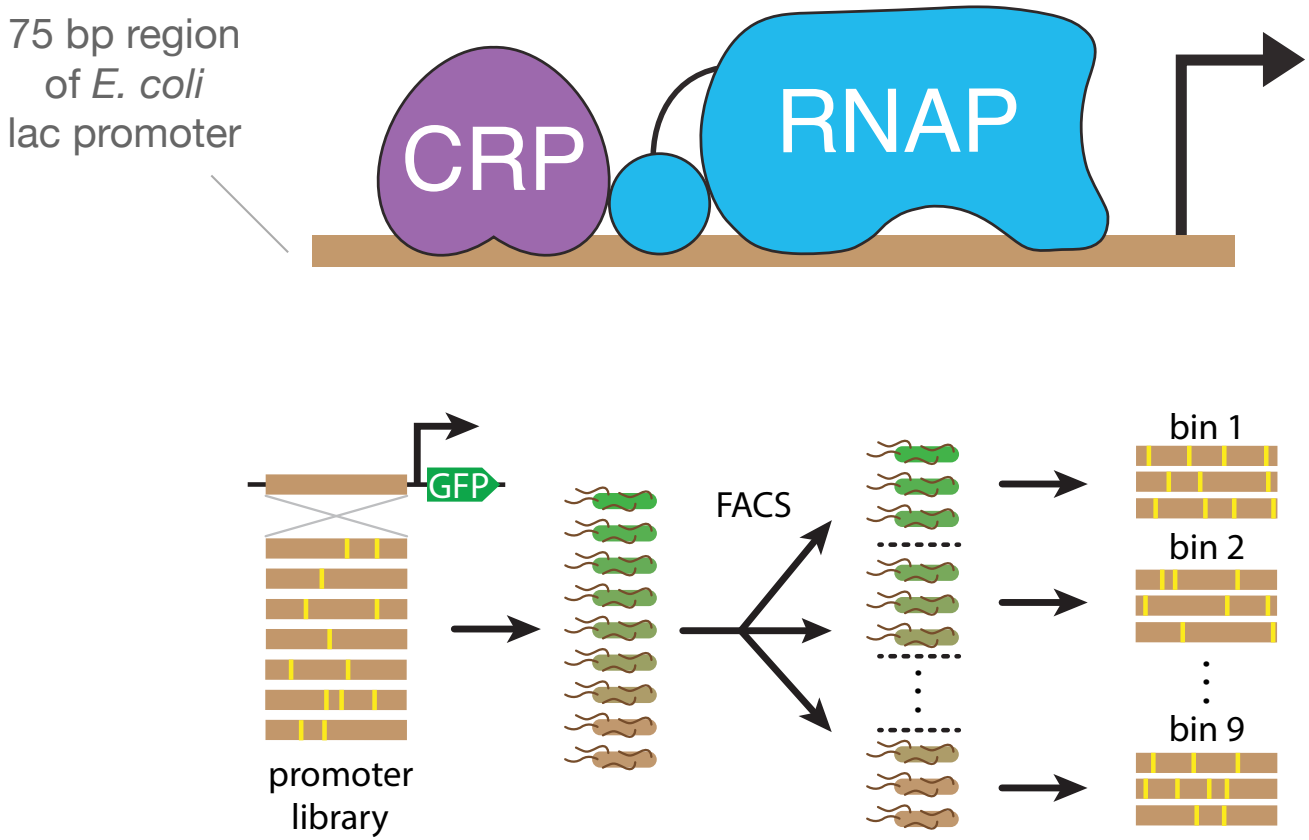
Goal: use high-throughput mutagenesis/ sequencing to systematically dissect the biophysical mechanisms of gene regulation.

Anand Murugan

Curtis Callan
(Princeton)

Edward Cox
(Princeton)

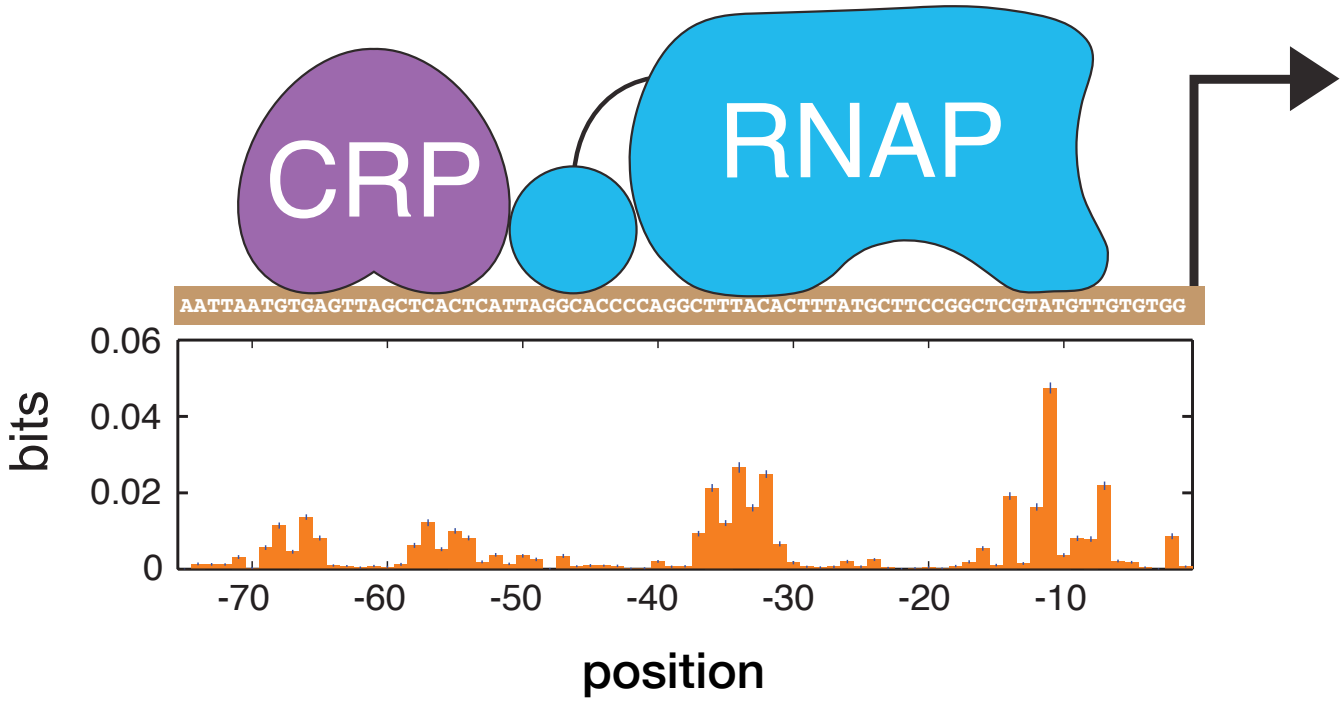
Sort-Seq is a massively parallel reporter assay (MPRA) capable of simultaneously measuring the activities of many thousands of gene regulatory sequences.



Simultaneous measurements for ~250,000 different promoters

Kinney et al., 2010, *PNAS*

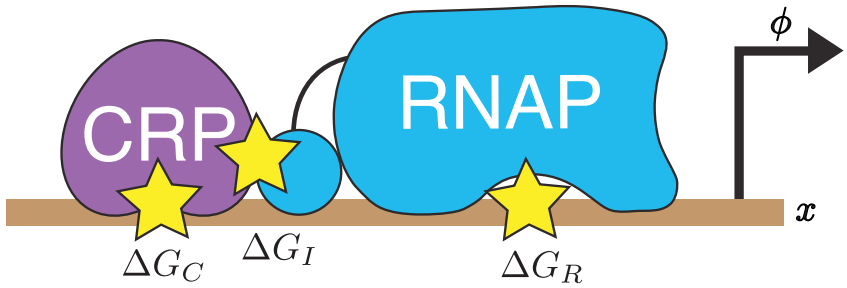
Sort-Seq enables the de novo identification of function binding sites via “information footprints”.



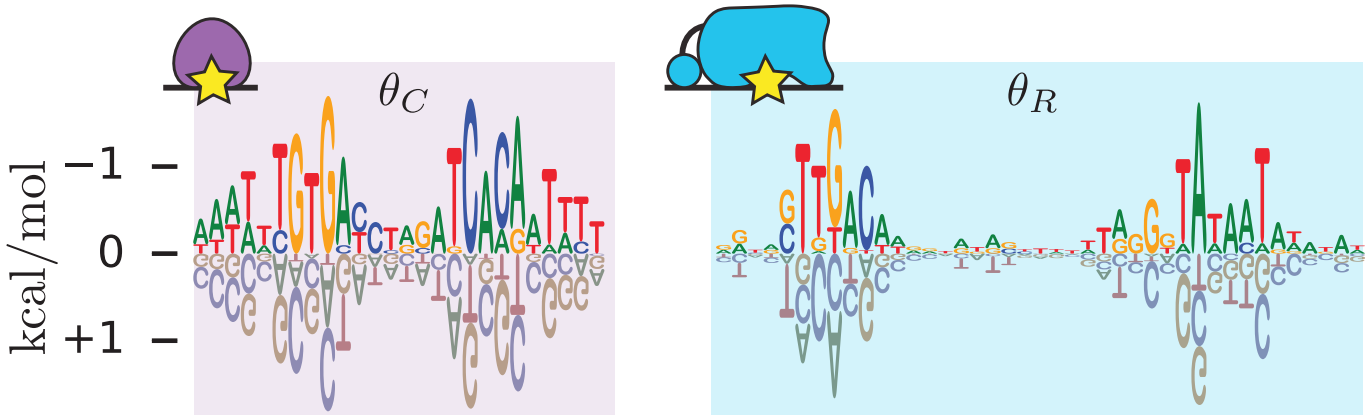
Kinney et al., 2010, *PNAS*

Fitting the parameters of biophysical models to Sort-seq data allows one to quantitatively define the biophysical mechanisms of gene regulation.

proposed model:



inferred parmeters:



$\Delta G_I = -3.3 \pm 0.4$ kcal/mol

Kinney et al., 2010, *PNAS*

Sort-Seq is just one example of a massively parallel assay, a.k.a. multiplex assay of variant effect (MAVE)

ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

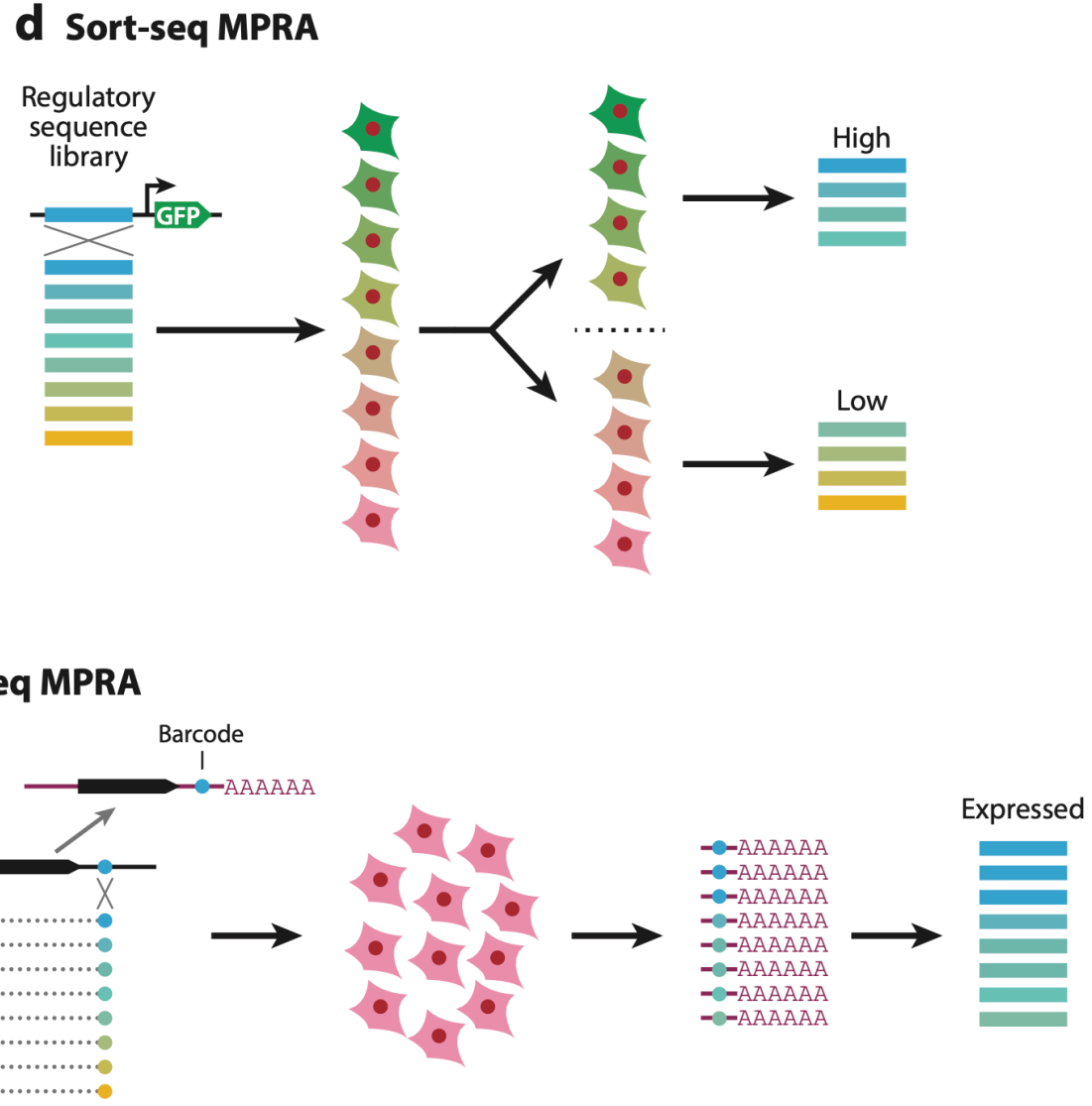
Massively Parallel Assays and Quantitative Sequence–Function Relationships

Justin B. Kinney and David M. McCandlish

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; email: jkinney@cshl.edu, mccandlish@cshl.edu

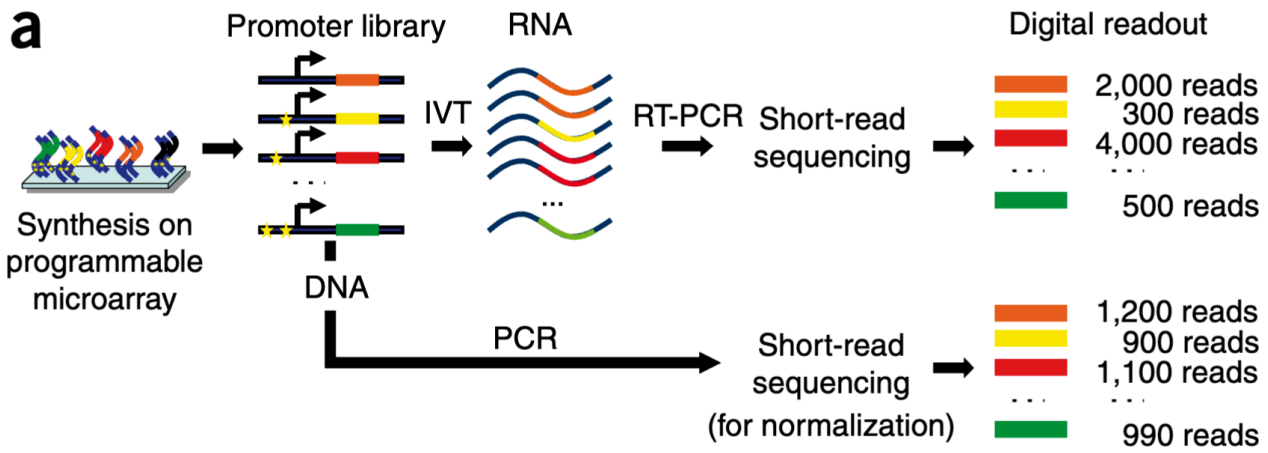
Kinney & McCandlish., 2019, *Annu Rev Genom Hum G*

Massively parallel reporter assays (MPRAs)

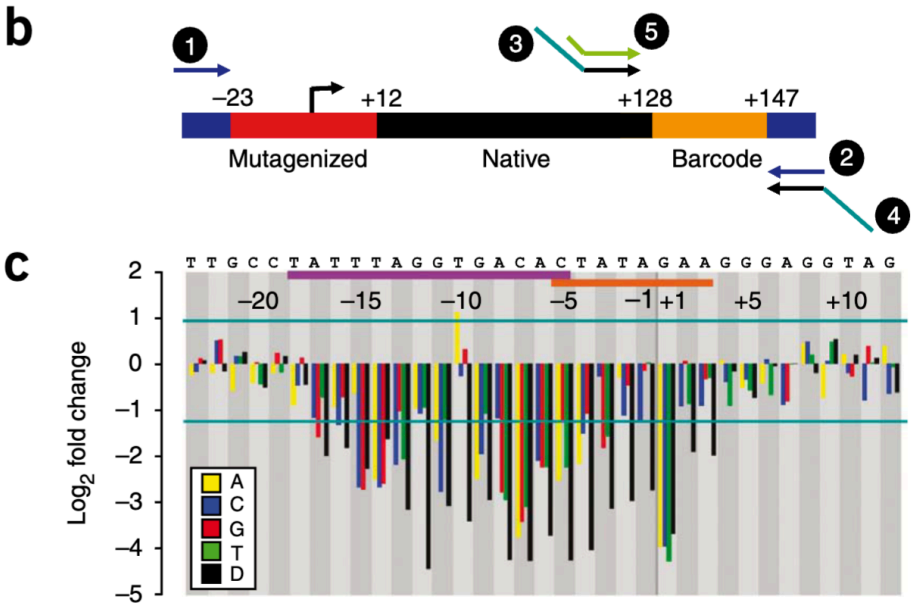


Kinney & McCandlish., 2019, *Annu Rev Genom Hum G*

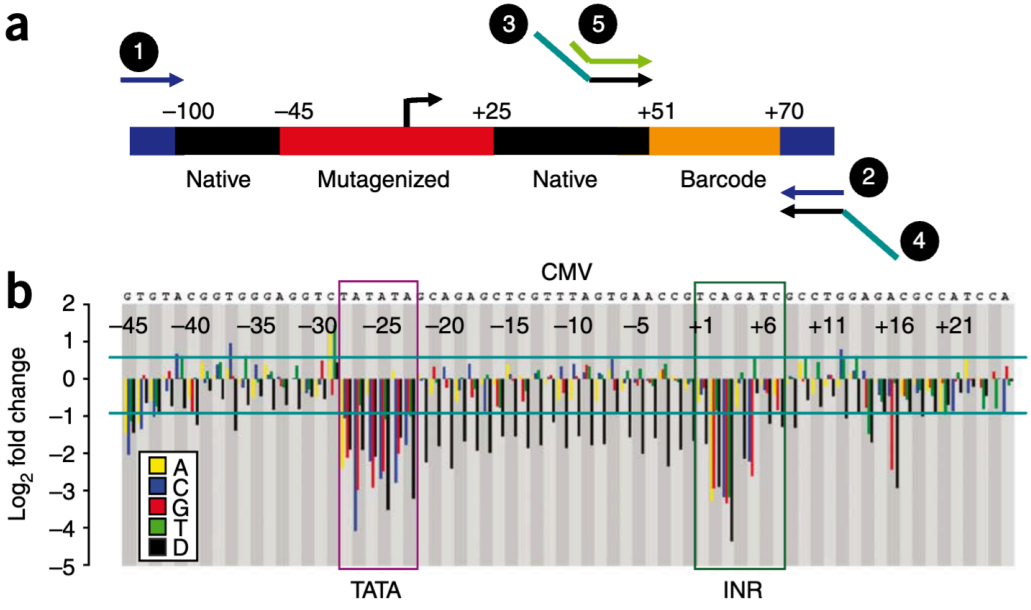
MPRAs on promoters in vitro
Patwardhan et al., 2009, Nat Biotech (Shendure Lab)



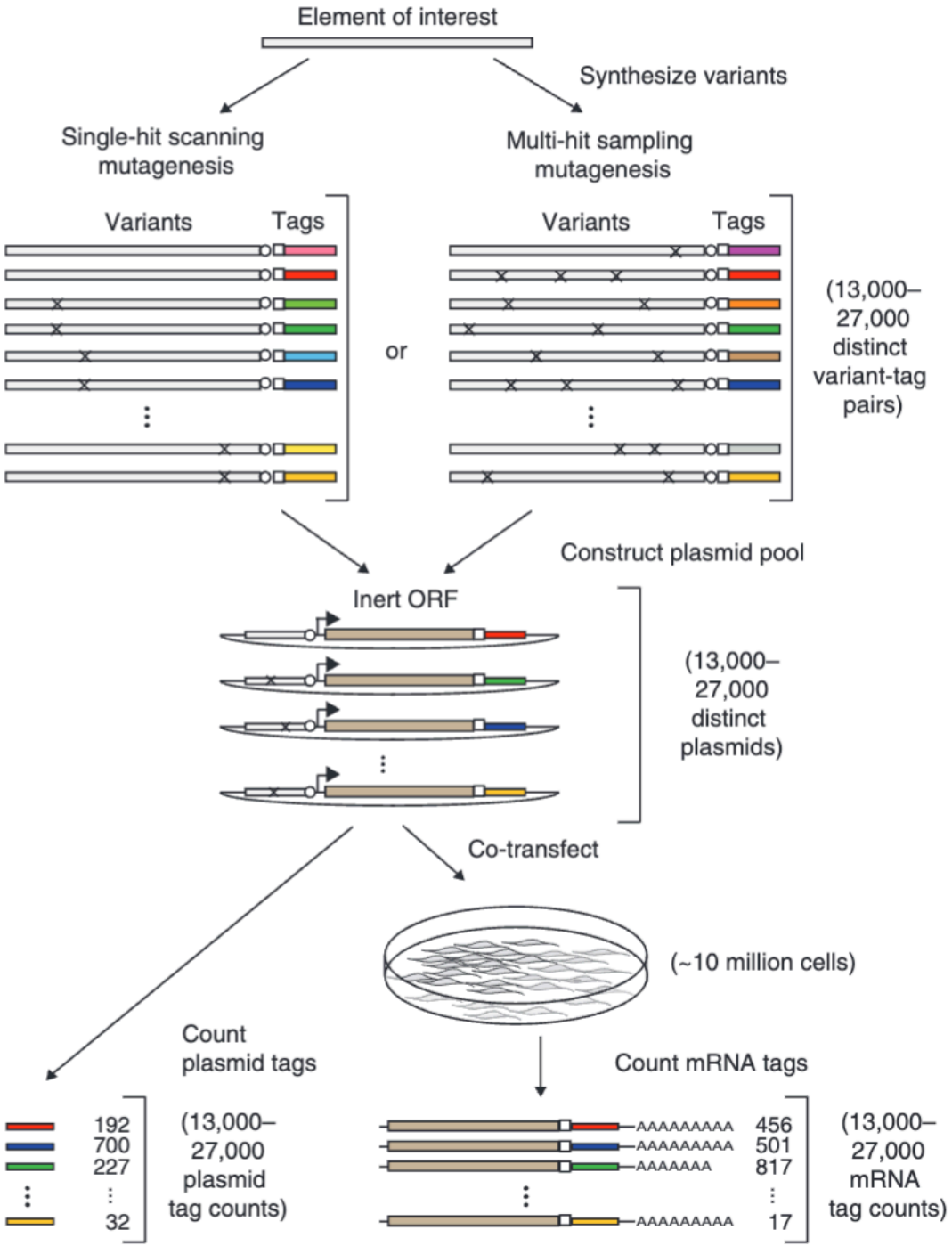
bacteriophage RNA polymerase



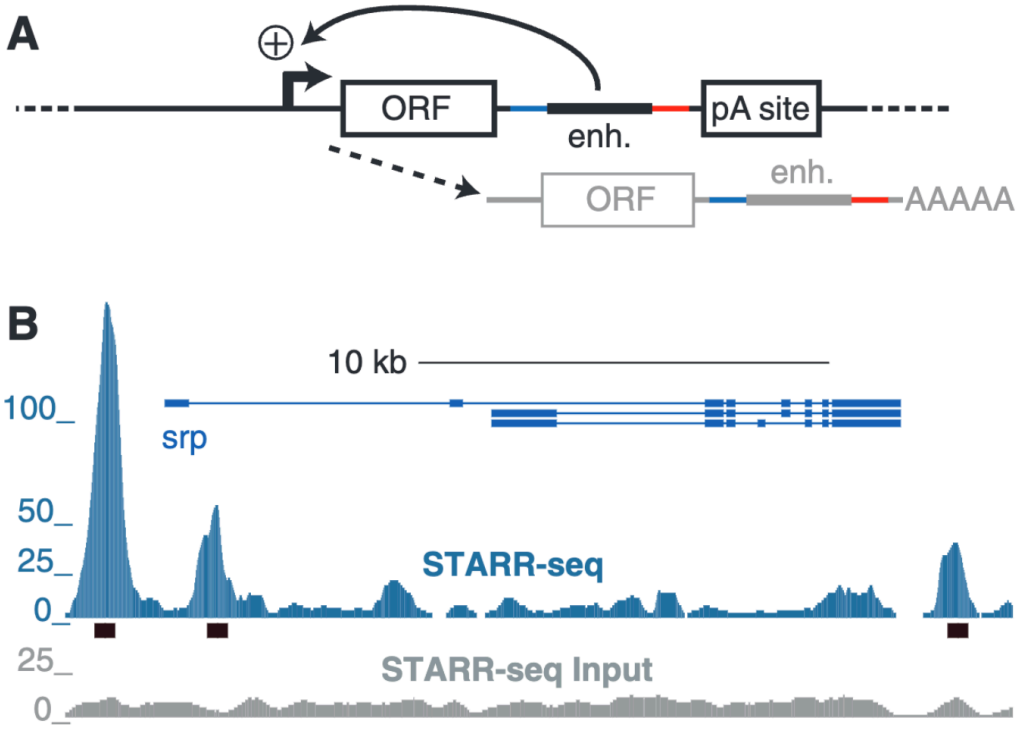
HeLa nuclear extracts

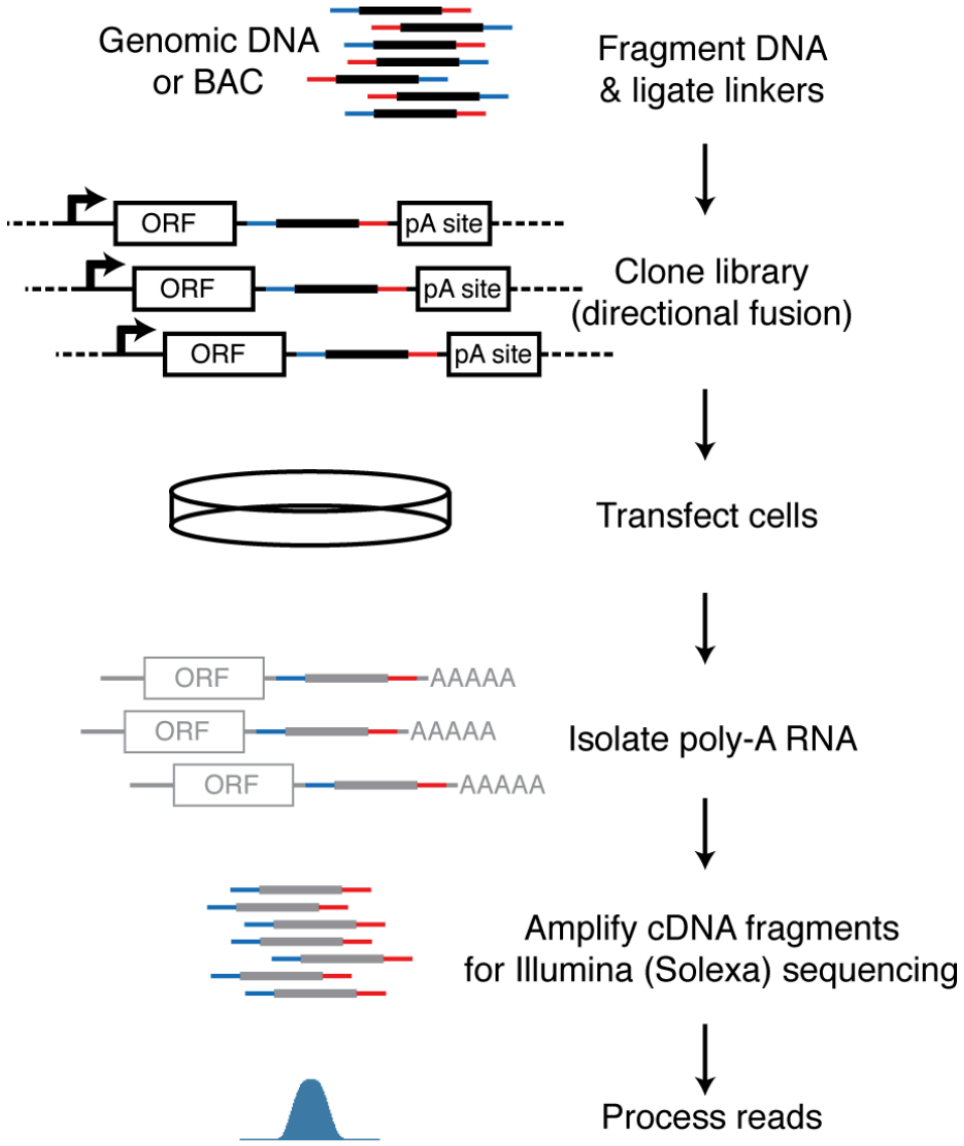


MPRAs on enhancers in human cells
Melnikov et al., 2012, Nat Biotech (Melnikov Lab)

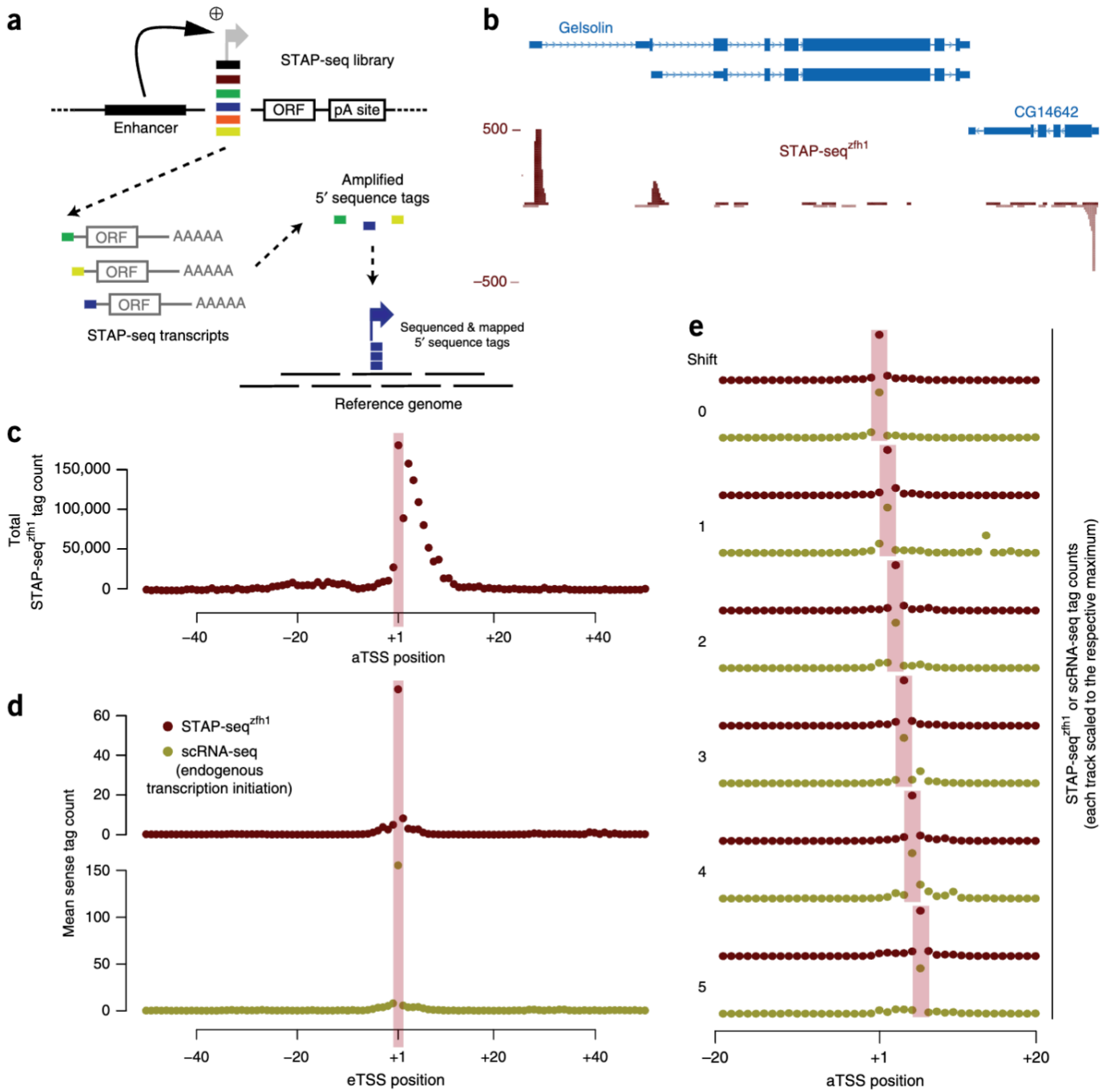


STARR-Seq
Arnold et al., 2014, Science (Stark Lab)

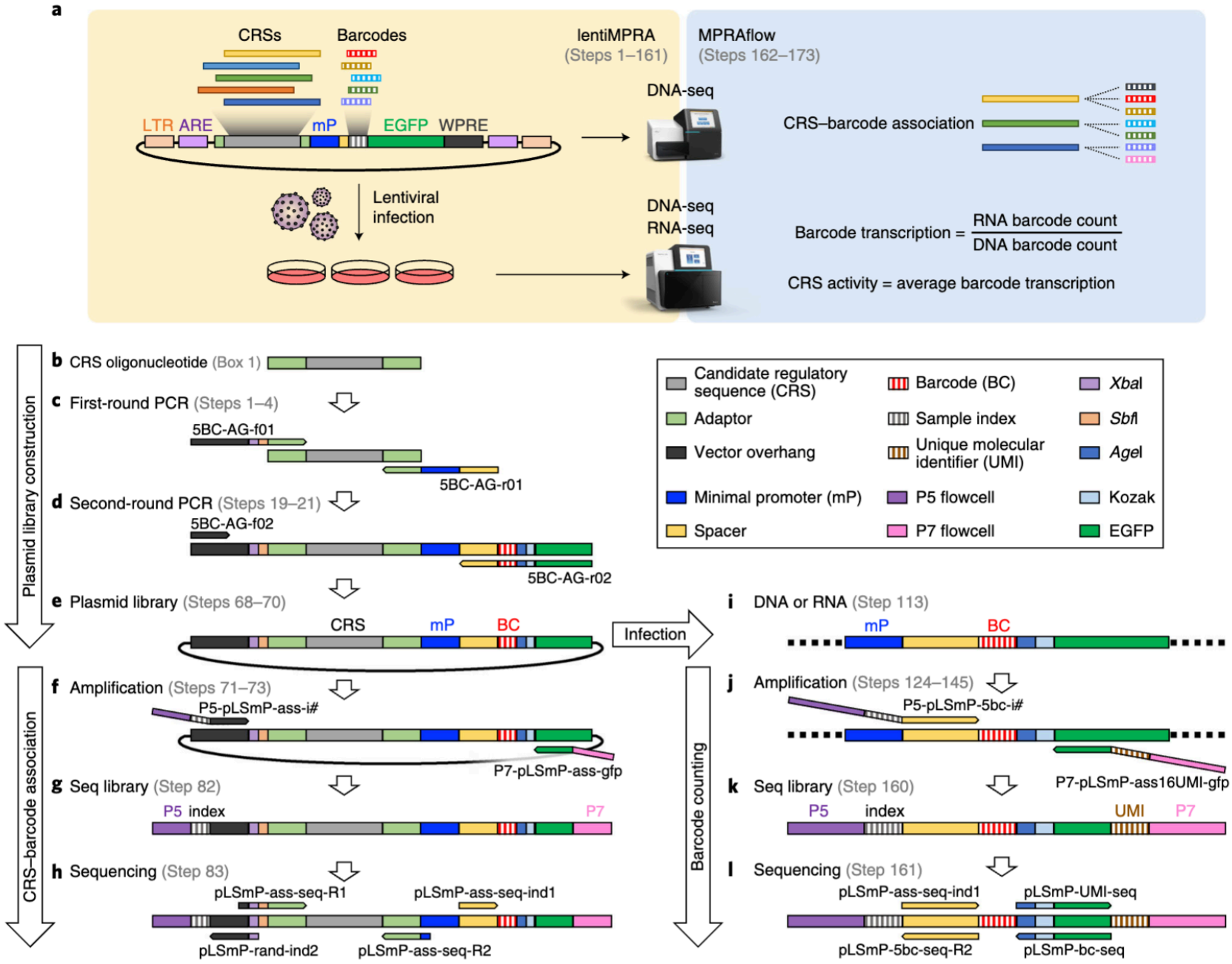




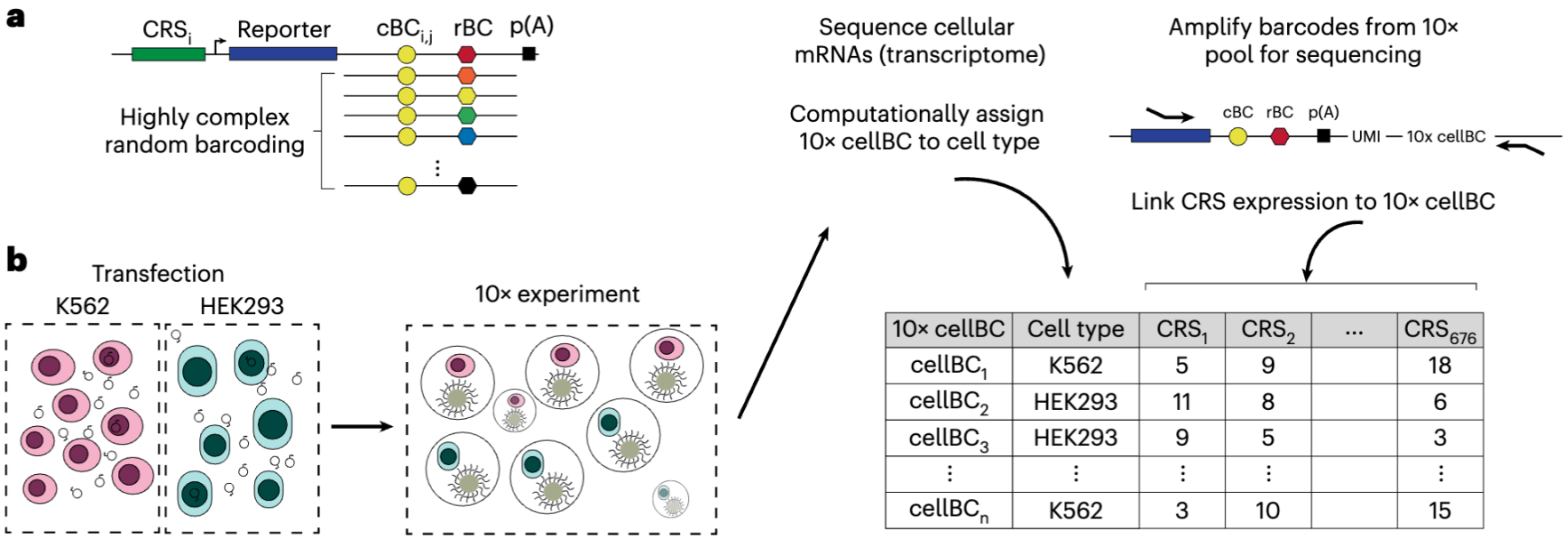
STAP-seq
Arnold et al., Nat Biotech, 2017 (Stark Lab)



LentiMPRA
Gordon et al., 2020, Nat Protocol (Ahituv Lab)



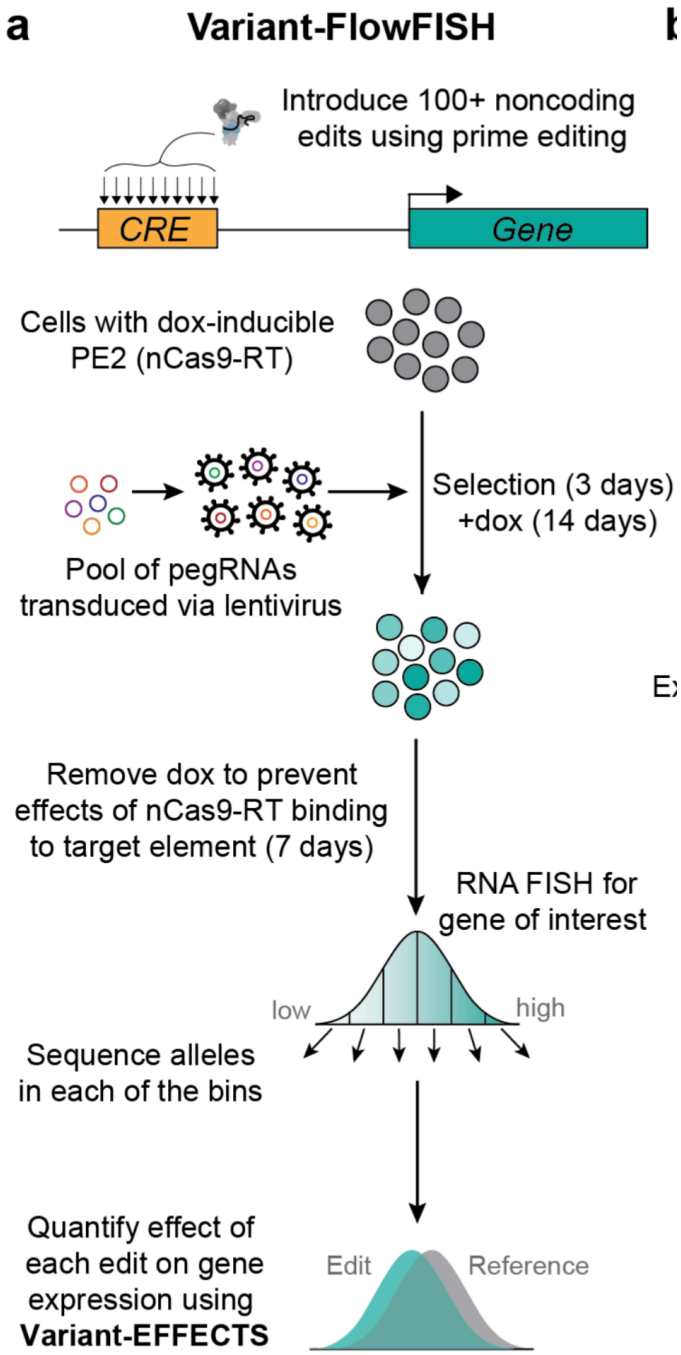
scMPRA
Zhao et al., 2023, Nat Genet (Cohen Lab)



10× cellBC	Cell type	CRS$_1$	CRS$_2$	...	CRS$_{676}$
cellBC$_1$	K562	5	9		18
cellBC$_2$	HEK293	11	8		6
cellBC$_3$	HEK293	9	5		3
⋮	⋮	⋮	⋮		⋮
cellBC$_n$	K562	3	10		15

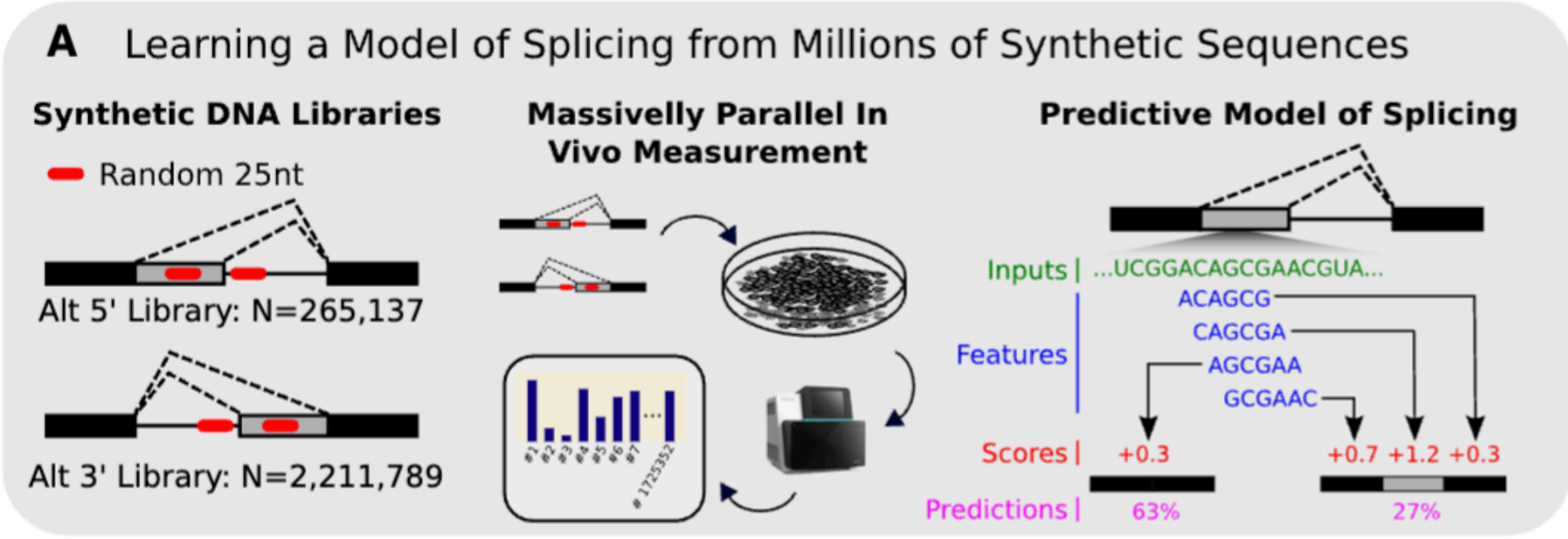
Fig. 1 | scMPRA measures CRS activity at single-cell resolution. a, Each CRS reporter construct is barcoded with a cBC that specifies the identity of the CRS, and a highly complex rBC. The complexity of the cBC–rBC pair ensures that the probability of identical plasmids being introduced into the same cell is extremely low. **b**, Experimental overview for scMPRA using the mixed-cell experiment as an example. K562 cells and HEK293 cells are transfected with the double-barcoded core promoter library. After 24 h, cells were collected and mixed for 10x scRNA-seq. Cell identities were obtained by sequencing the transcriptome, and single-cell expression from CRSs was obtained by quantifying the barcodes. The cell identity and CRS activities (as measured by the cBC–rBC abundances) were linked by the shared 10x cell barcodes.

Variant FlowFISH
Bartyn et al., 2023, bioRxiv (Engritz Lab)

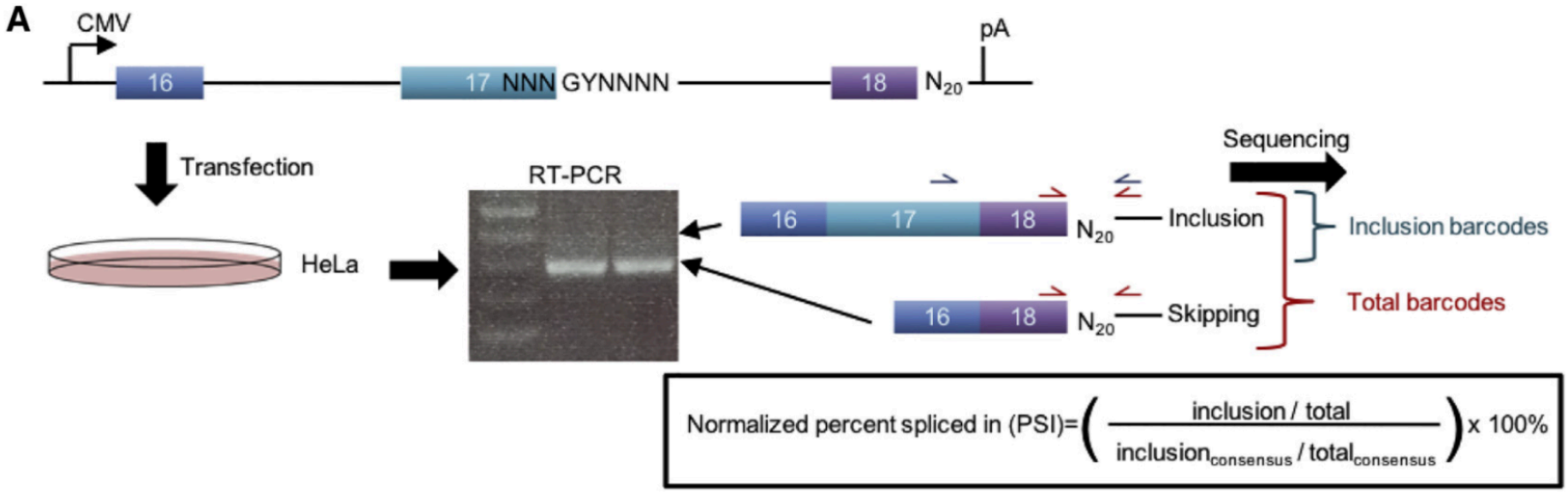


MPRAs for other aspects of mRNA processing

Massively parallel splicing assay
Rosenberg et al., 2015, Cell (Seelig Lab)



Massively parallel splicing assay
Wong et al., 2018, Mol Cell (Kinney & Krainer Labs)



MPRA via saturation genome editing
Findlay et al., 2014, Nature (Shendure Lab)

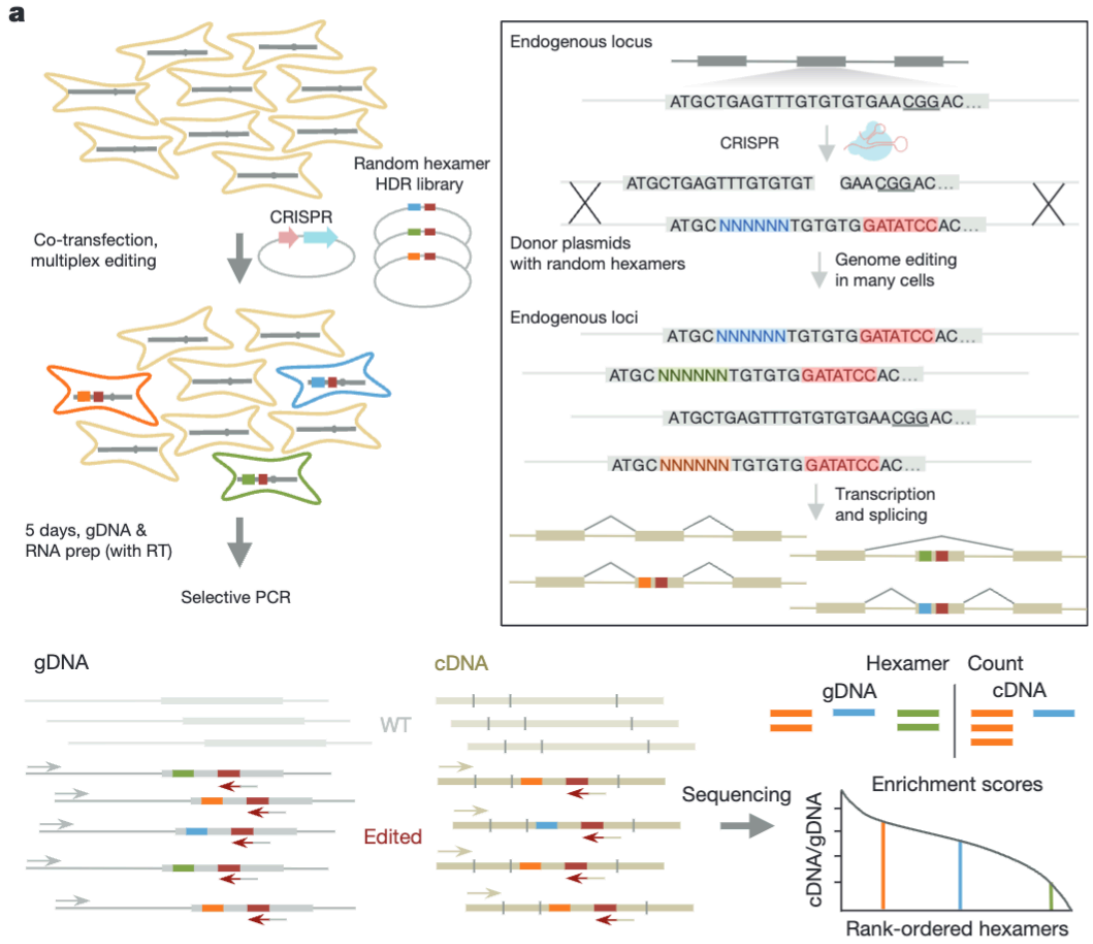
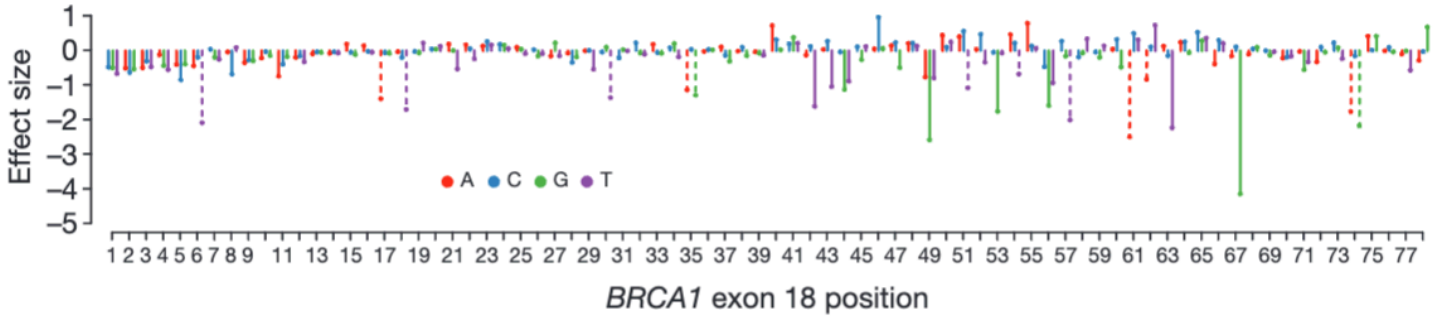
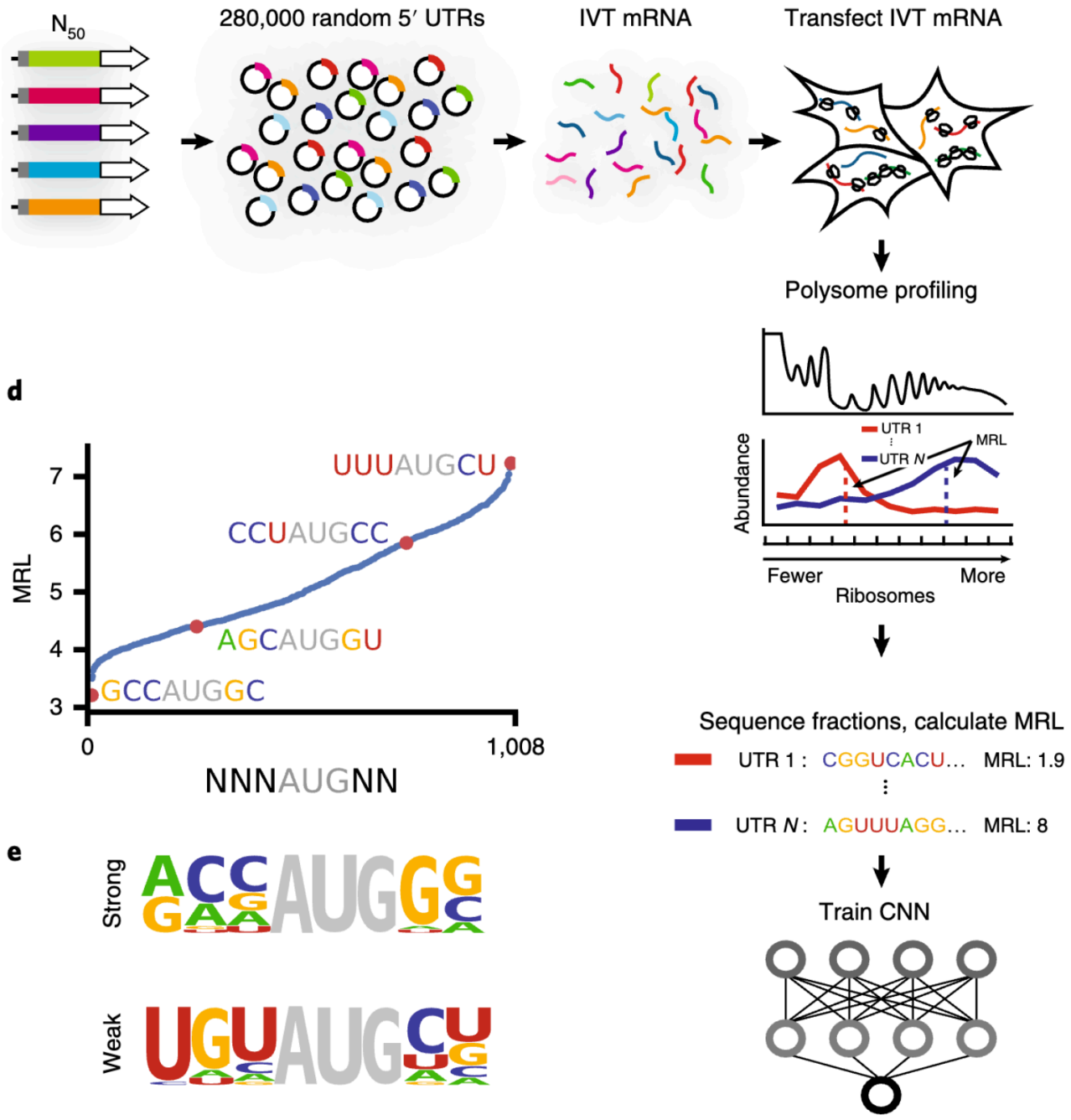


Figure 1 | Saturation genome editing and multiplex functional analysis of a hexamer region influencing *BRCA1* splicing. **a**, Experimental schematic. Cultured cells were co-transfected with a single Cas9-sgRNA construct (CRISPR) and a complex homology-directed repair (HDR) library containing an edited exon that harbours a random hexamer (blue, green, orange) and a fixed selective PCR site (red). CRISPR-induced cutting stimulated homologous recombination with the HDR library, inserting mutant exons into the genomes of many cells. At five days post-transfection, cells were harvested for gDNA and RNA. After reverse transcription, selective PCR was performed followed by sequencing of gDNA- and cDNA-derived amplicons. Hexamer enrichment scores were calculated by dividing cDNA counts by gDNA counts.

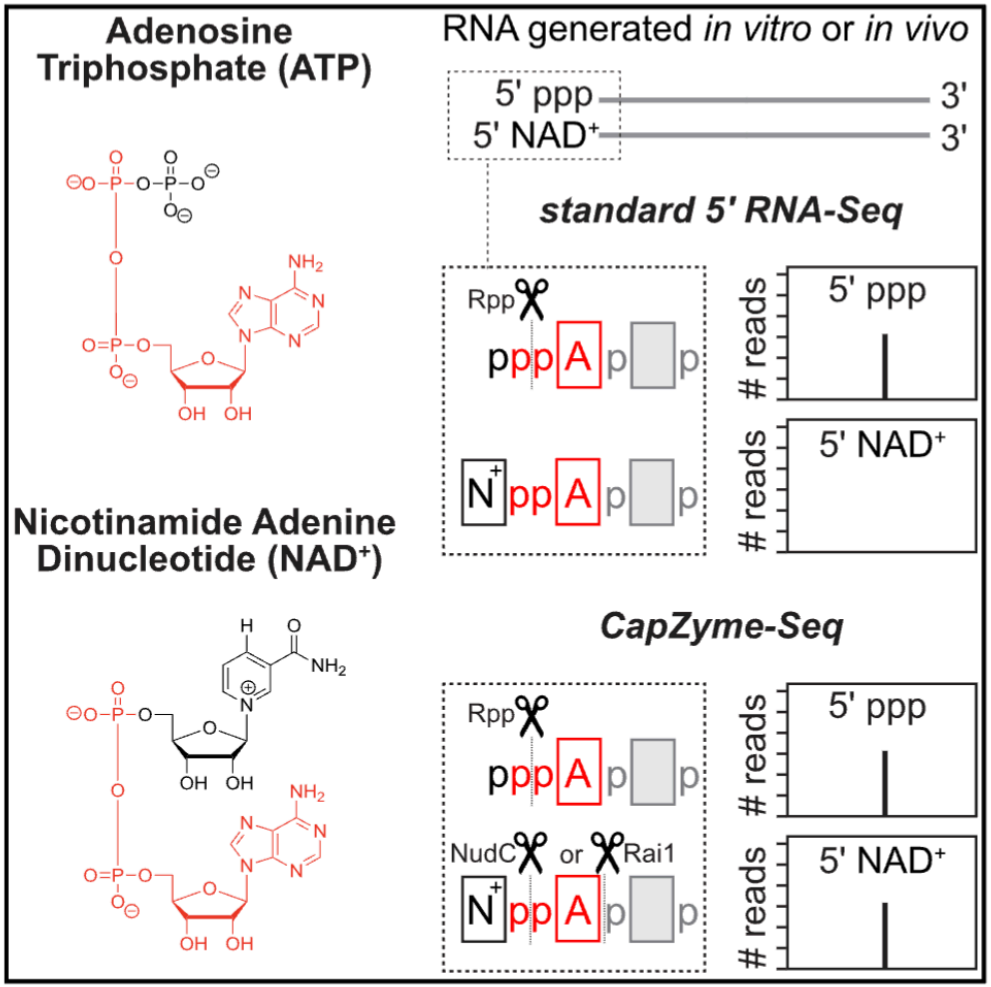


5’ UTR MPRA by ribosome profiling
Sample et al., 2019, Nat Biotech (Seelig Lab)



CapZyme-Seq
Vvedenskaya et al., 2018, Mol Cell (Nickels & Ebright Labs)

Graphical Abstract



Authors

Irina O. Vvedenskaya, Jeremy G. Bird, Yuanchao Zhang, ..., Deanne M. Taylor, Richard H. Ebright, Bryce E. Nickels

Correspondence

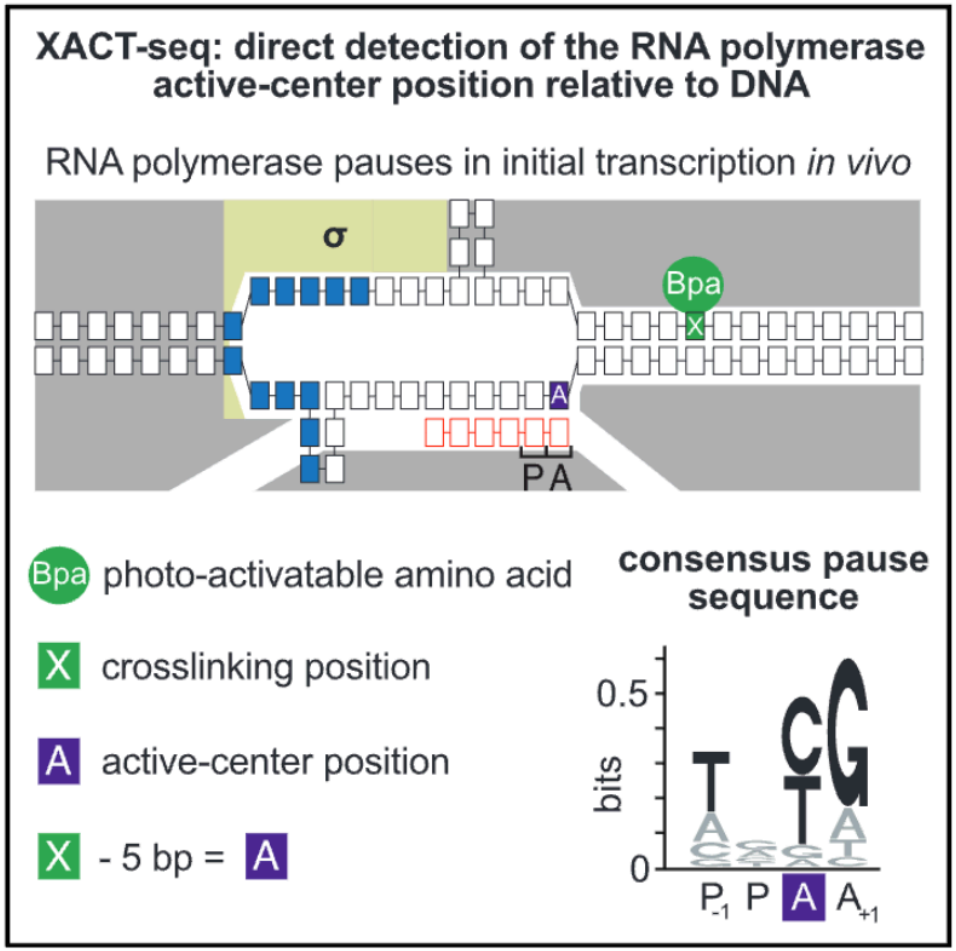
ebright@waksman.rutgers.edu (R.H.E.), bnickels@waksman.rutgers.edu (B.E.N.)

In Brief

Vvedenskaya et al. report a high-throughput-sequencing-based technology that employs NAD+-decapping enzymes to detect and quantify NAD+-capped RNA. Analysis of NAD+ capping for ~16,000 promoter sequences defines a consensus promoter sequence for NAD+ capping.

XACT-Seq
Winkleman et al., 2020, Mol Cell (Nickels & Ebright Labs)

Graphical Abstract



Authors

Jared T. Winkelman,
Chirangini Pukhrambam,
Irina O. Vvedenskaya, ..., Premal Shah,
Richard H. Ebright, Bryce E. Nickels

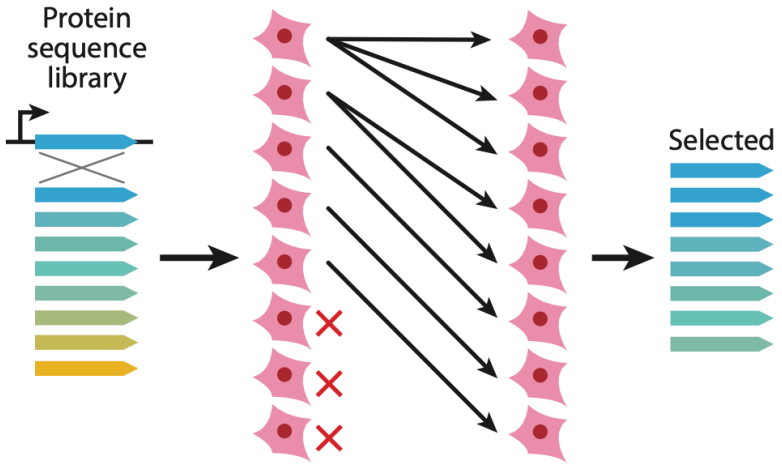
Correspondence

ebright@waksman.rutgers.edu (R.H.E.),
bnickels@waksman.rutgers.edu (B.E.N.)

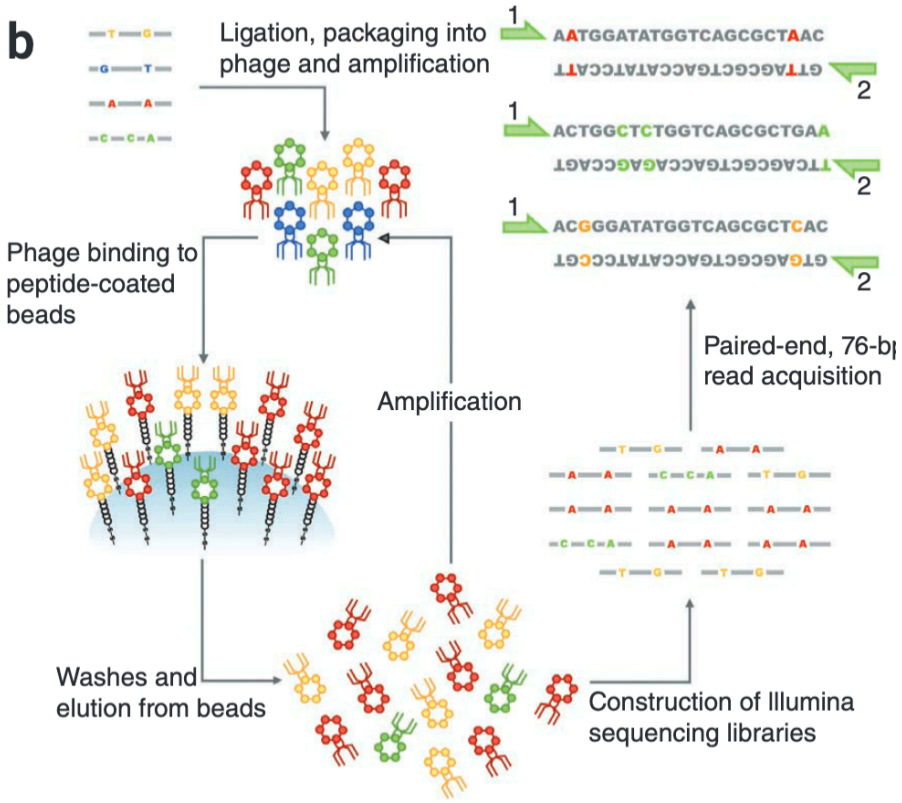
In Brief

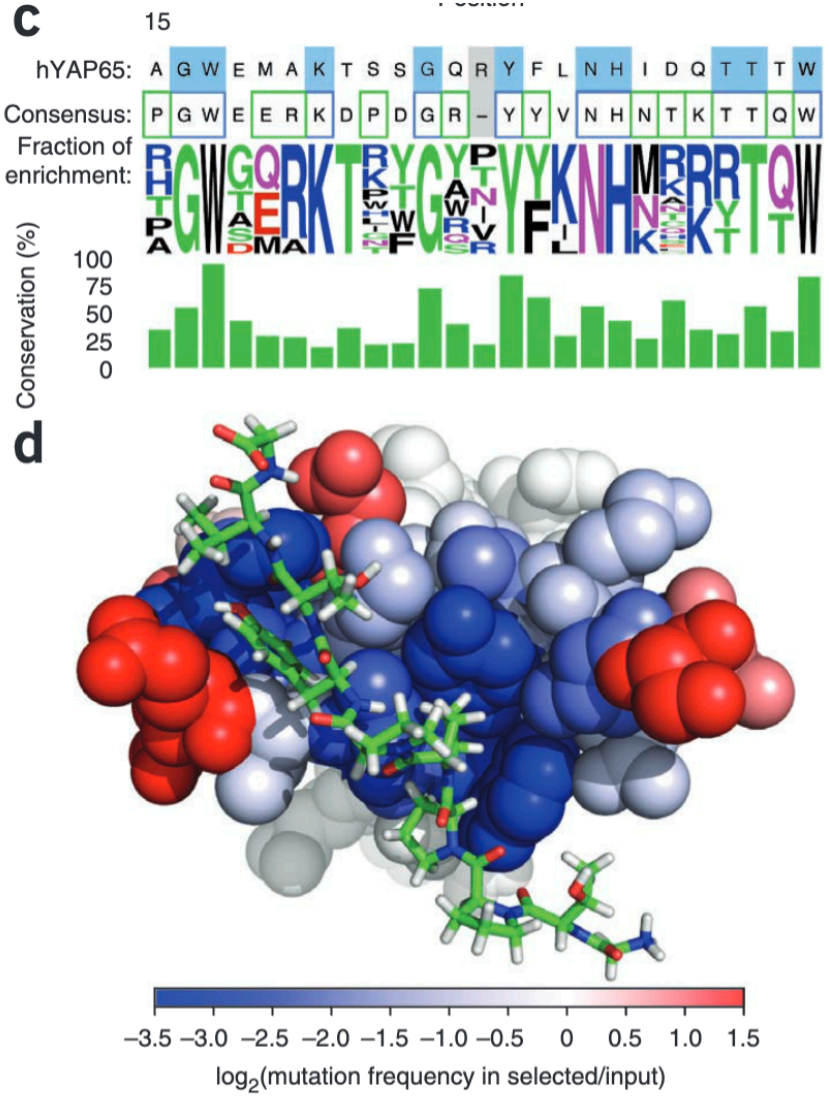
Winkelman et al. report a protein-DNA photocrosslinking method that provides single-nucleotide-resolution readout of RNA polymerase active-center position relative to DNA and enables analysis of initial-transcription pausing. Analysis of 4^{11} (~4,000,000) promoter sequences defines positional determinants and sequence determinants for initial-transcription pausing by bacterial RNA polymerase *in vitro* and *in vivo*.

DMS approaches



Deep Mutational Scanning (phage display)
Fowler et al., 2010, Nat Meth (Fields Lab)





Deep Mutational Scanning (selective growth)
Hietpas et al., 2011, PNAS (Bolon Lab)

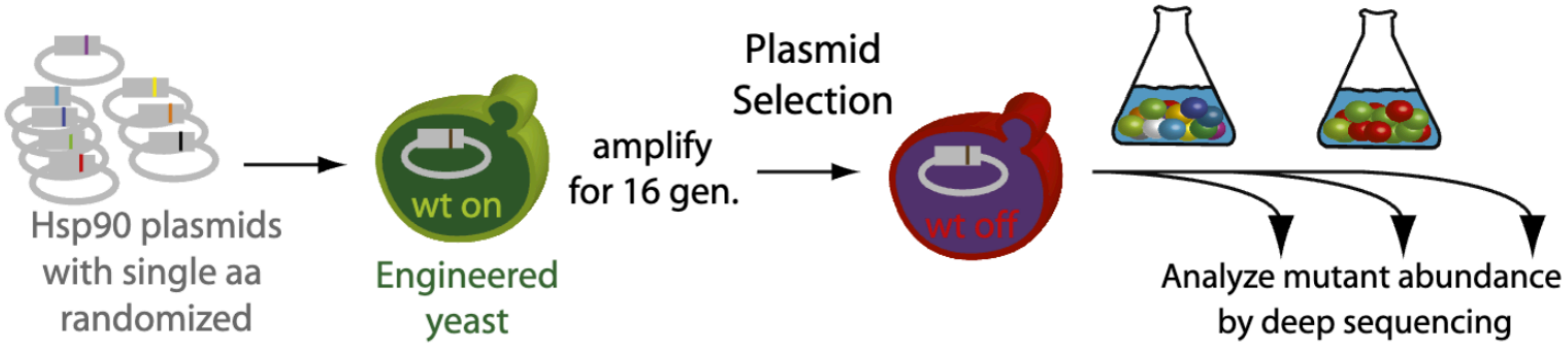
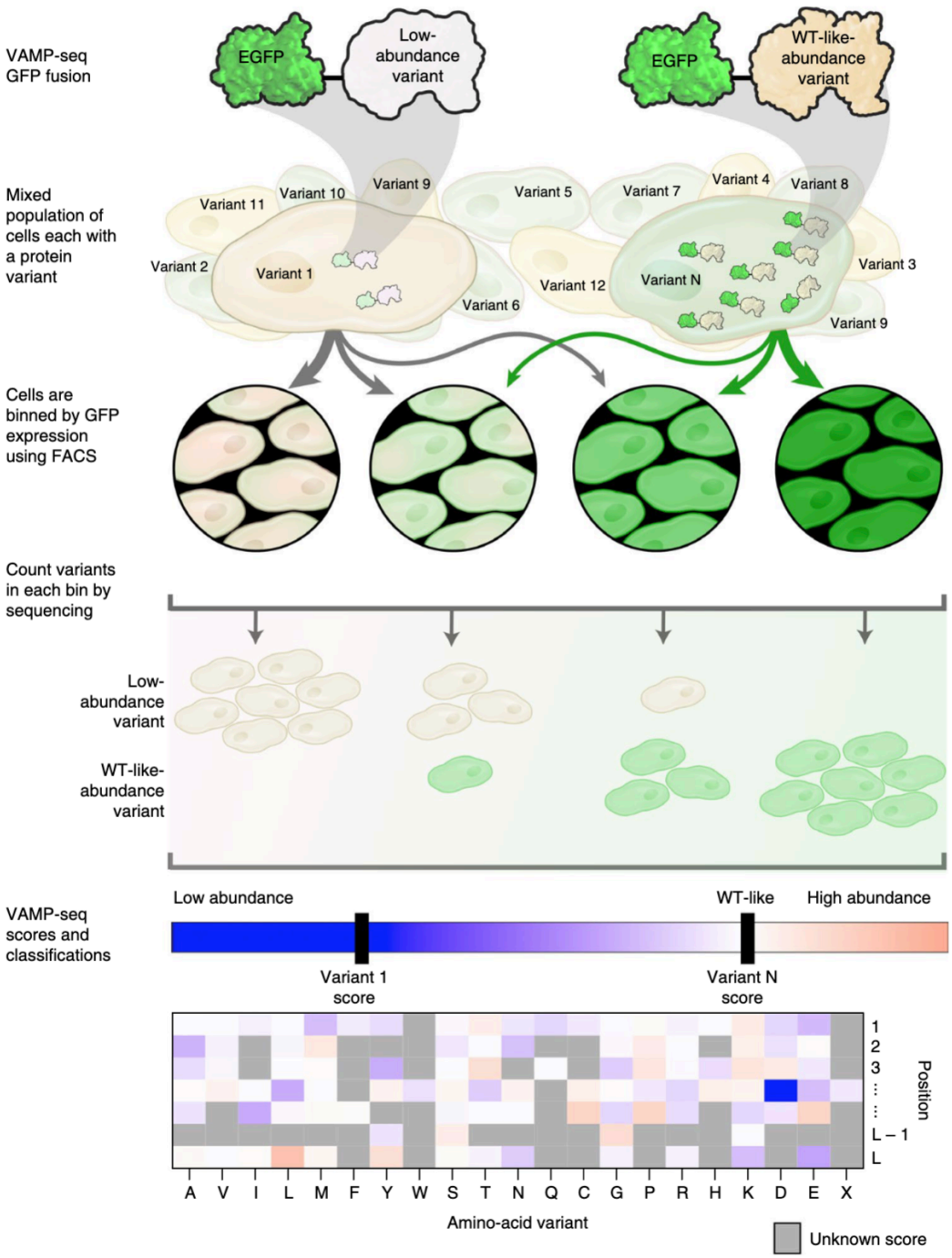
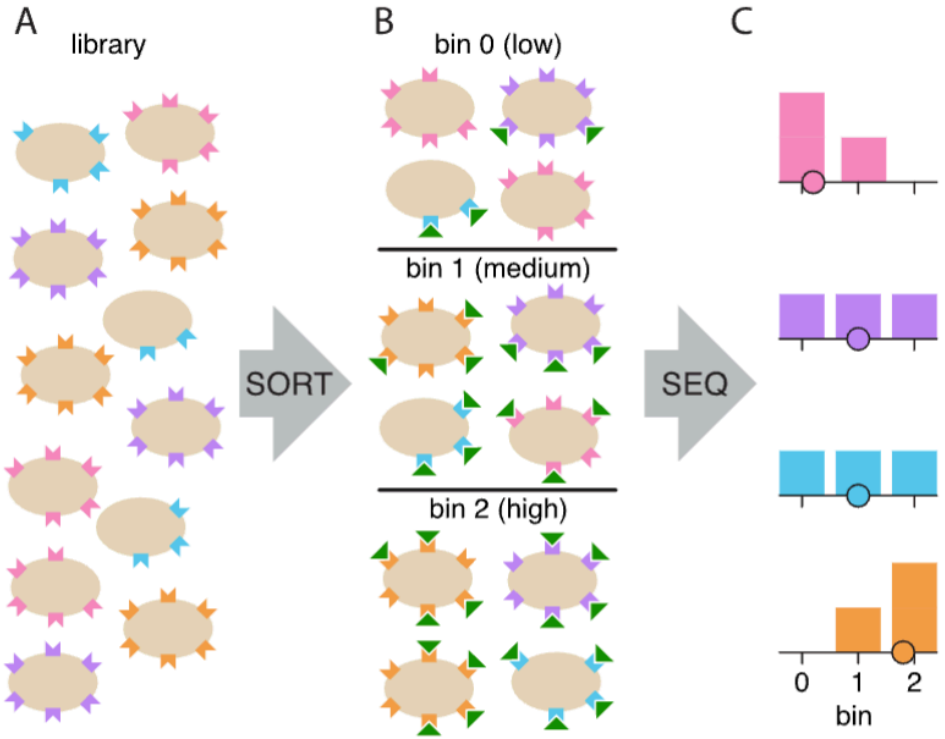


Fig. 1. EMPIRIC approach to experimentally determine fitness landscapes. Randomized individual codon libraries are introduced into a host cell whose only other copy of the gene is regulatable. The fitness of each individual codon mutation is determined by measuring its abundance in the mixed culture as a function of time under selective conditions.

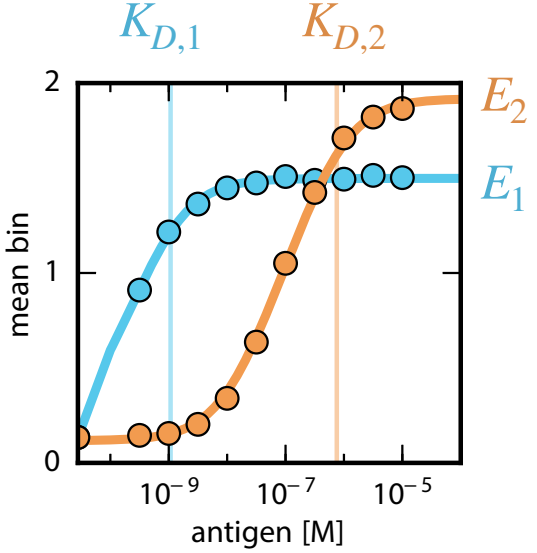
VAMP-seq
Matreyek et al., 2018, Nat Genet (Fowler Lab)



Tite-Seq
Adams et al., 2016, eLife (Kinney Lab)

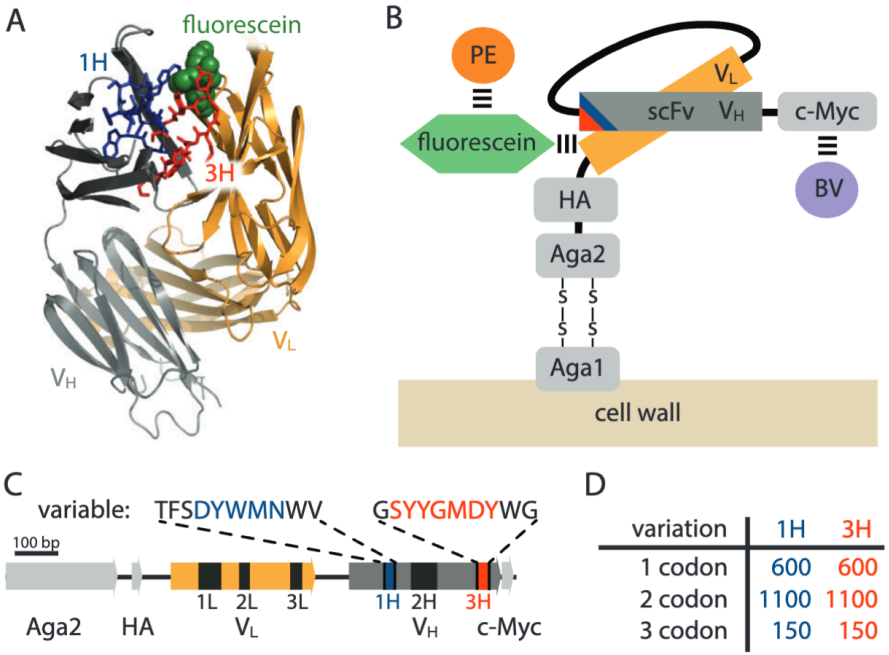


absolute affinity measurements (dissociation constants)



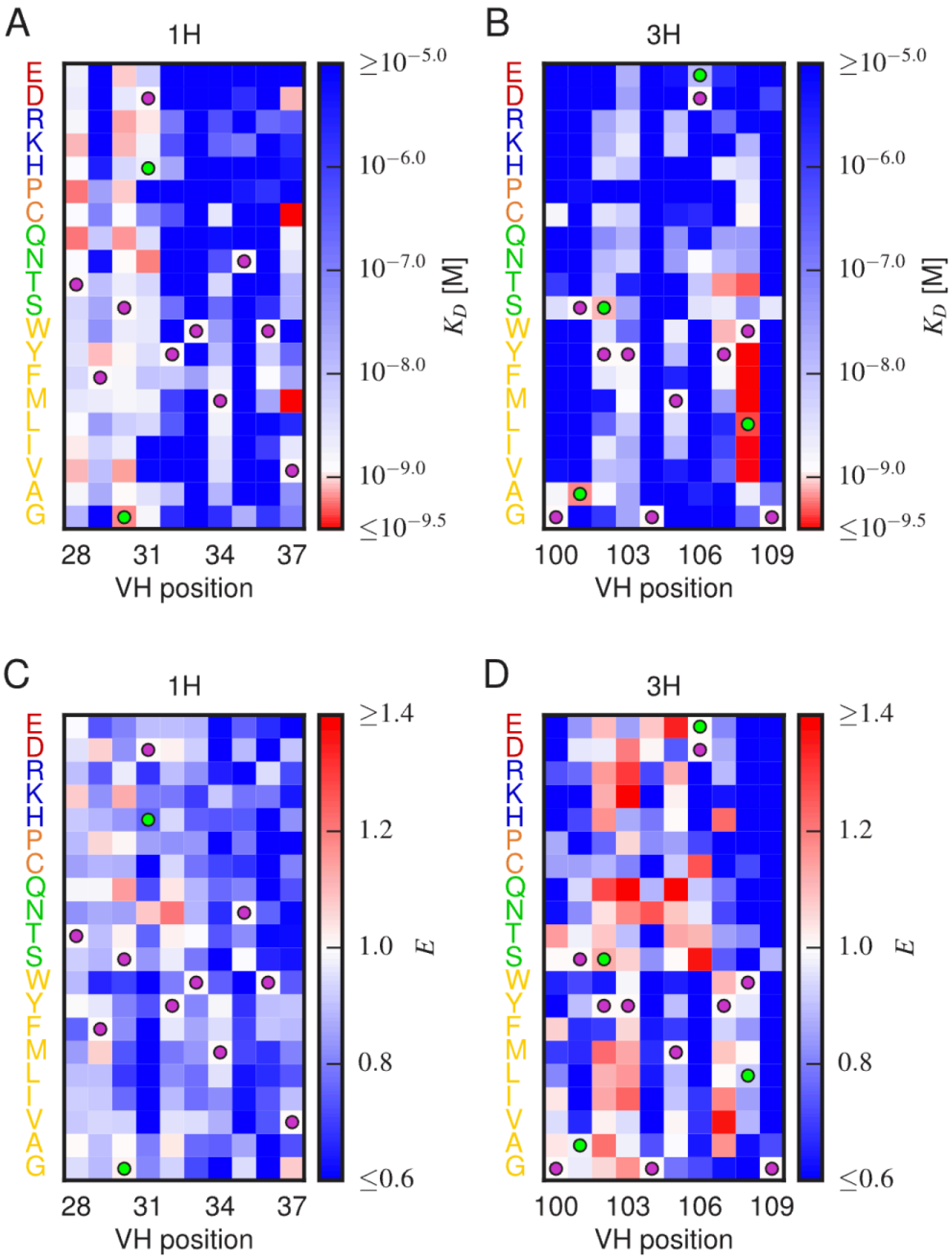
relative expression measurements

Tite-Seq
Adams et al., 2016, eLife (Kinney Lab)

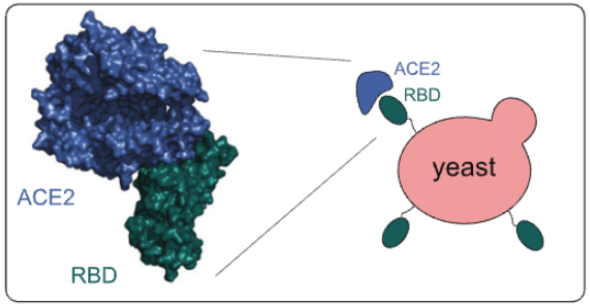


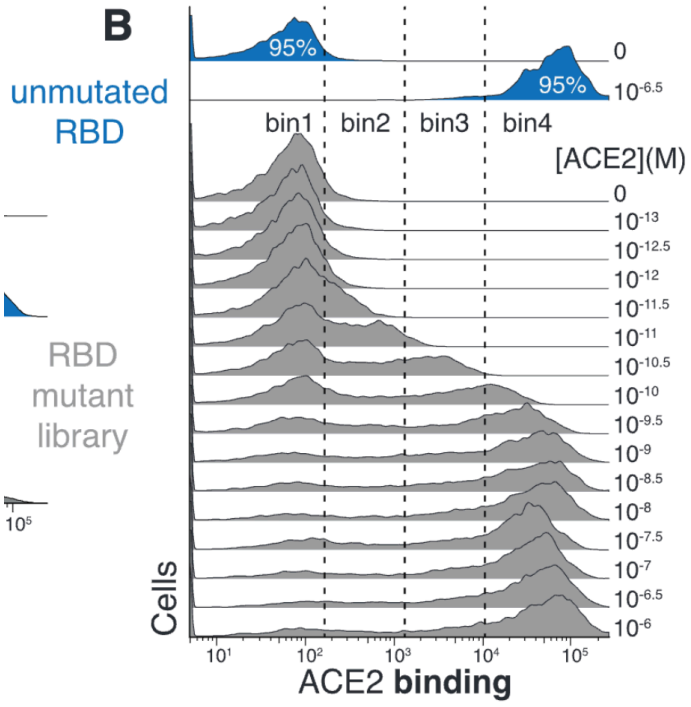
D

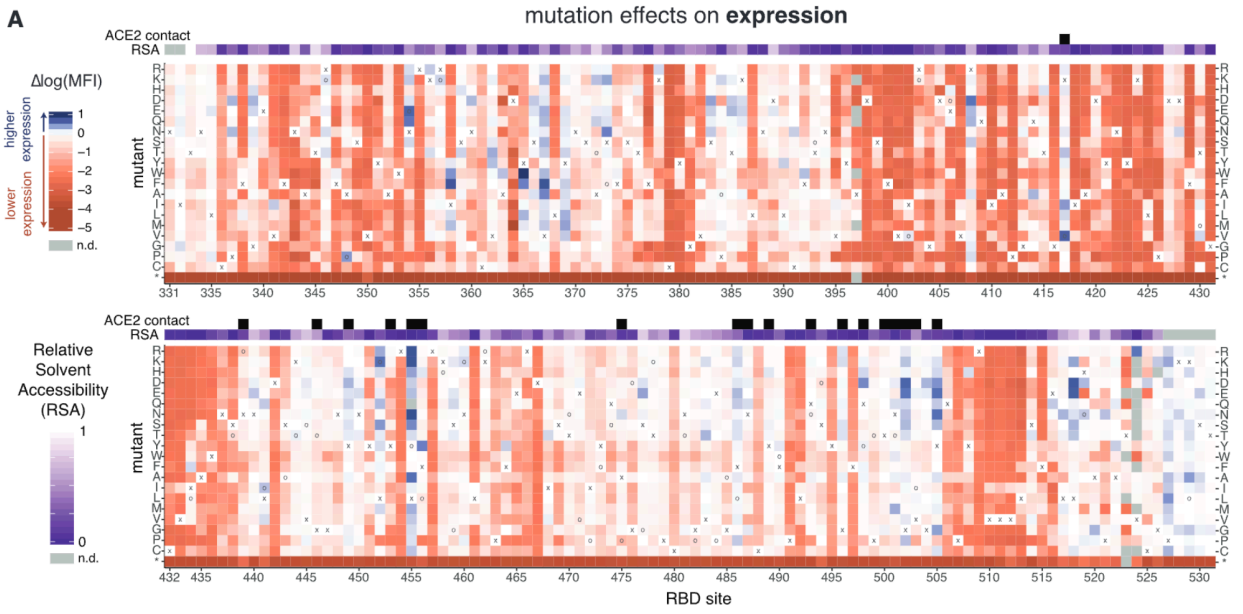
variation	1H	3H
1 codon	600	600
2 codon	1100	1100
3 codon	150	150

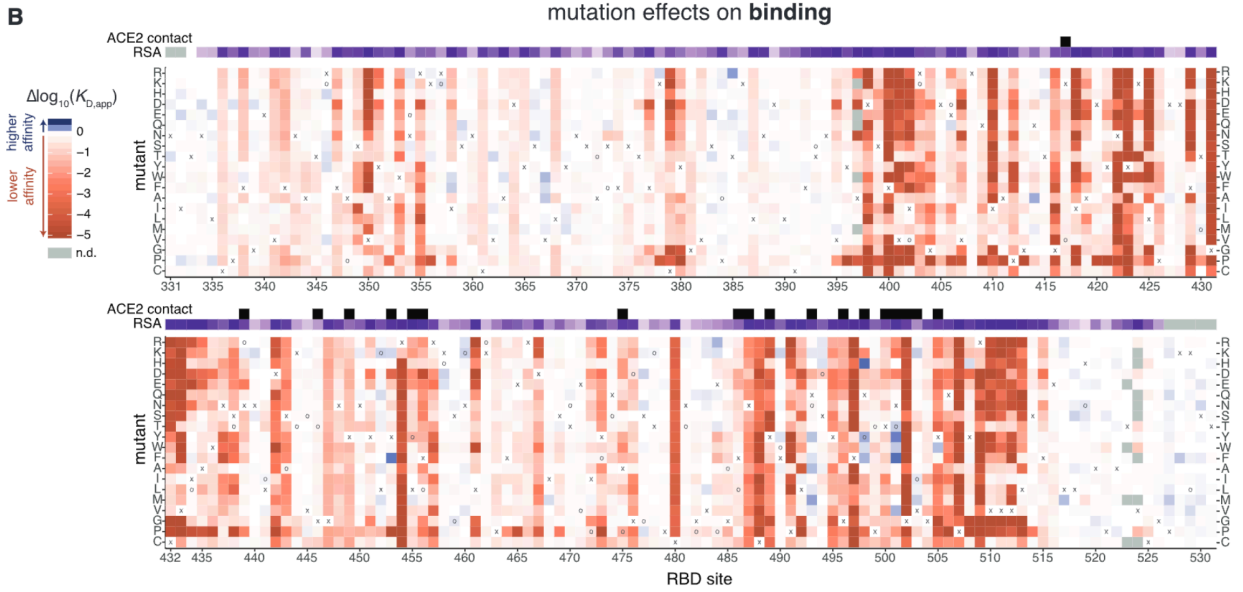


Tite-Seq applied to SARS-CoV-2
Starr et al., 2020, Cell (Bloom Lab)

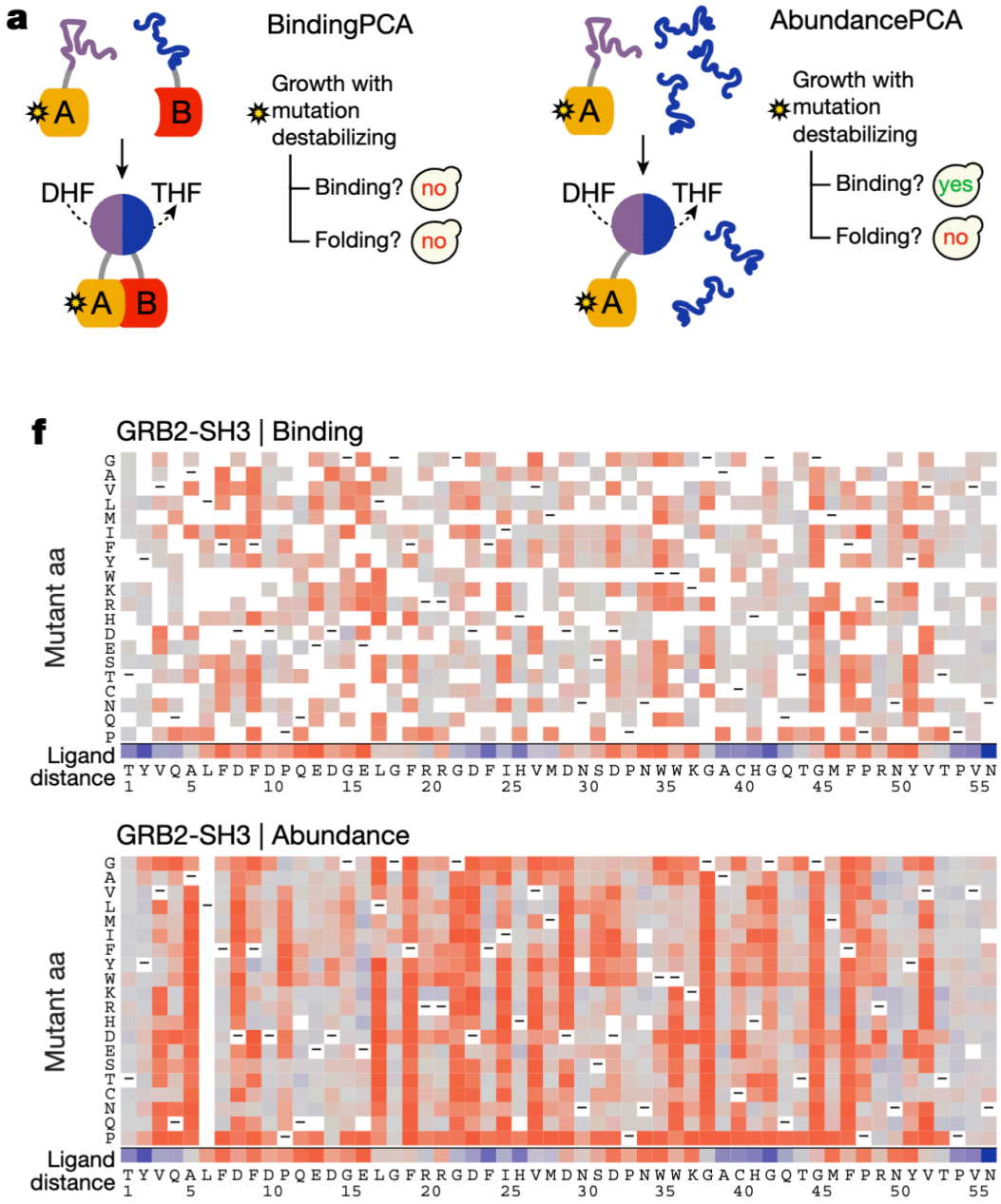






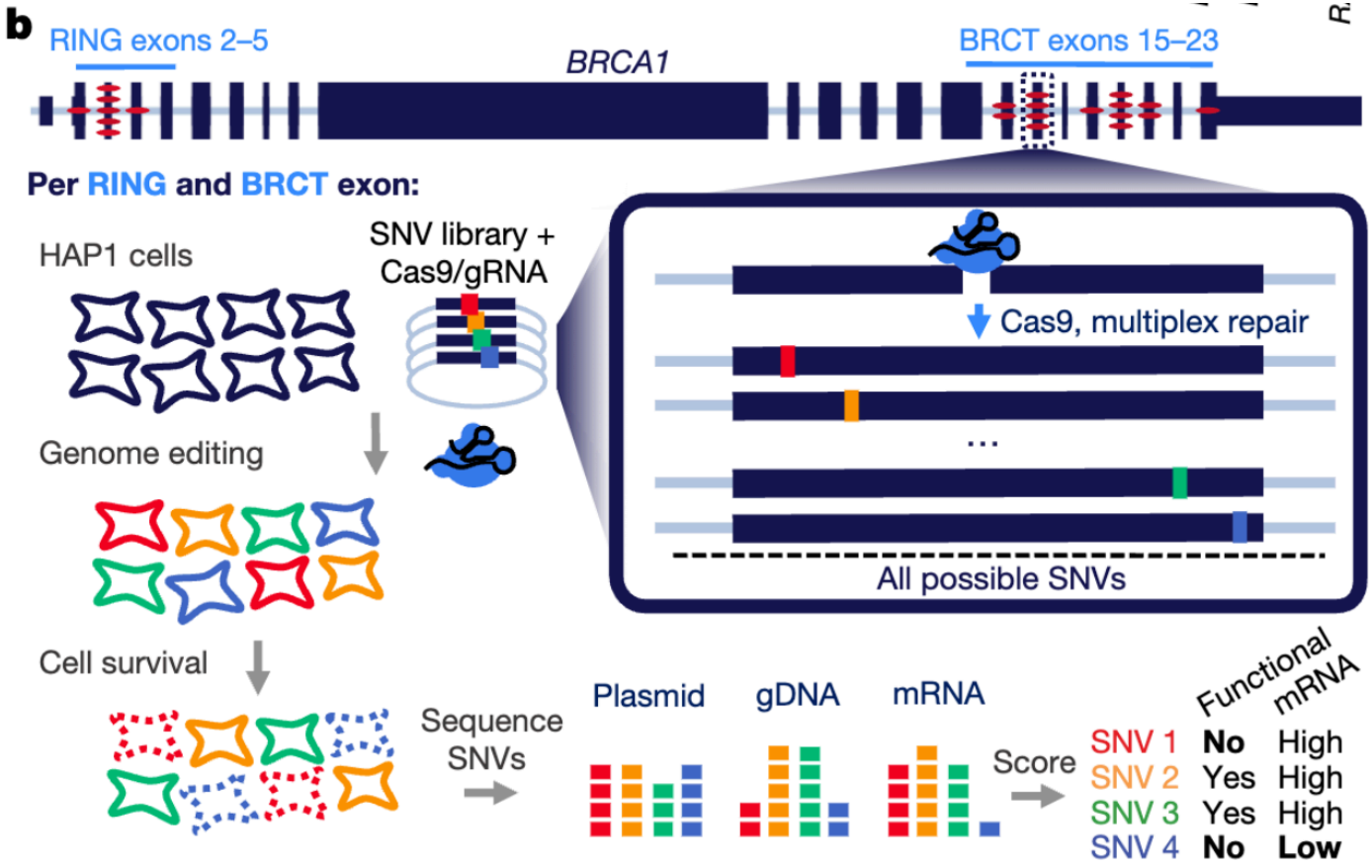


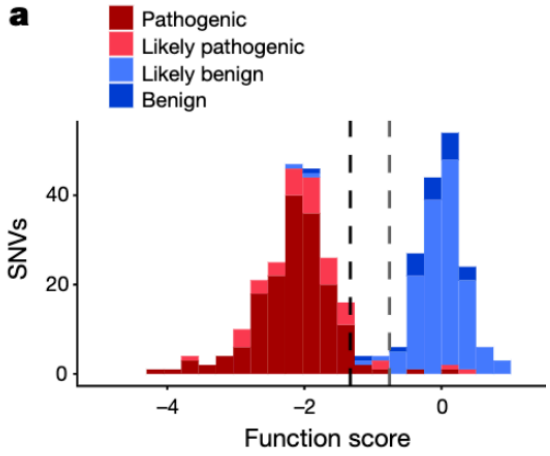
ddPCA (doubledeep protein fragment complementation)
Faure et al., 2022, Nature (Lehner Lab)

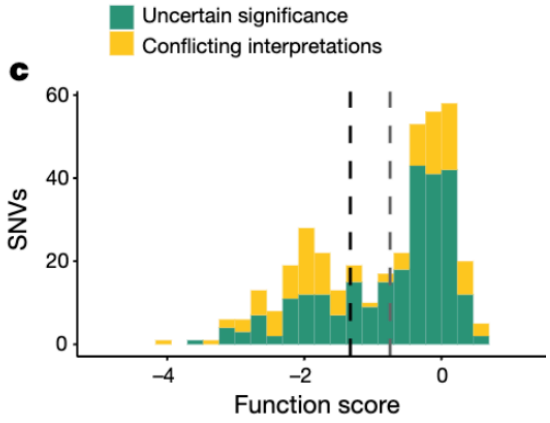


We therefore developed a strategy that uses two separate selection assays based on protein fragment complementation (PCA) to quantify the effects of mutations on both the abundance of a protein and its binding to an interaction partner (Fig. 1a). As perturbations to probe the potential for allosteric regulation, we use mutations; these are a convenient method to introduce diverse changes in chemistry at all sites in a protein[20,23]. In the first assay, 'BindingPCA', the binding between two proteins is quantified by fusing them to different fragments of a reporter enzyme, dihydrofolate reductase (DHFR). Interaction between the proteins brings the DHFR fragments into close proximity, allowing them to form a functional enzyme whose activity as measured by cellular growth in selective conditions is proportional to the intracellular concentration of the protein complex[24]. In the second assay, 'AbundancePCA', only one protein is expressed and fused to a DHFR fragment with the other DHFR fragment being highly expressed. Functional DHFR is now reconstituted by random encounters and growth is proportional to the intracellular concentration of the first protein over more than 3 orders of magnitude, as validated by applying the assay to more than 2,000 yeast proteins[25]. We refer to the combination of these two assays as 'doubledeepPCA' (ddPCA), a high-throughput method that quantifies the effects of mutations on both the abundance of a protein and its binding to one or more interaction partners. ddPCA builds on and extends previous work using PCA to probe the effects of mutations on protein binding and stability[26,27].

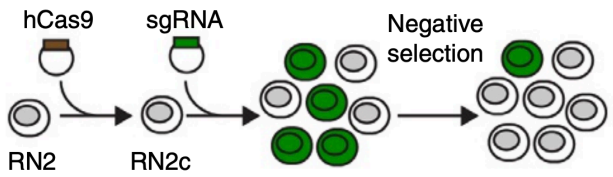
DMS via saturation genome editing
Findlay et al., 2018, Nature (Shendure Lab)

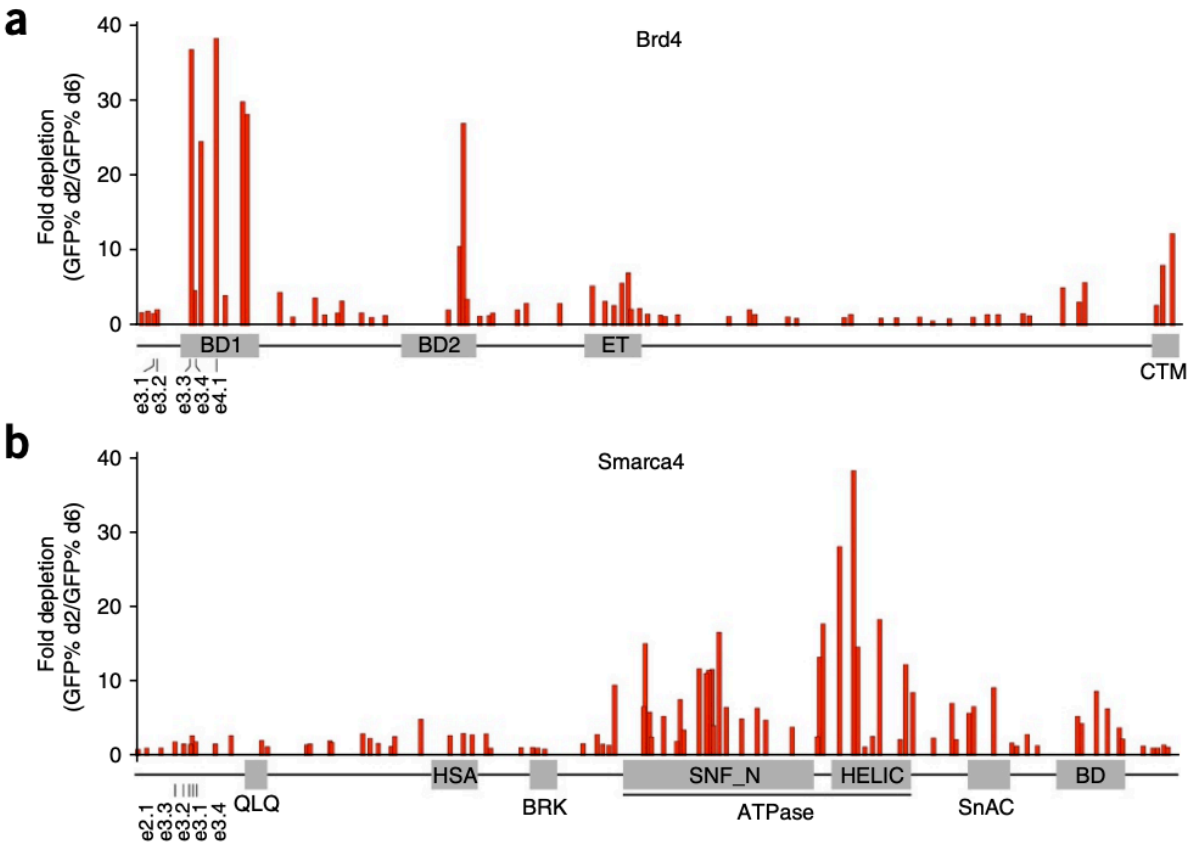




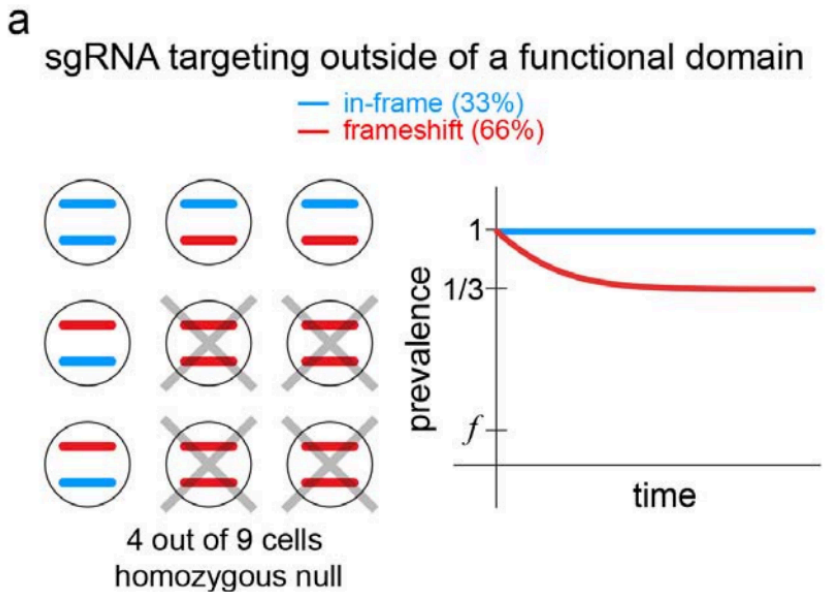


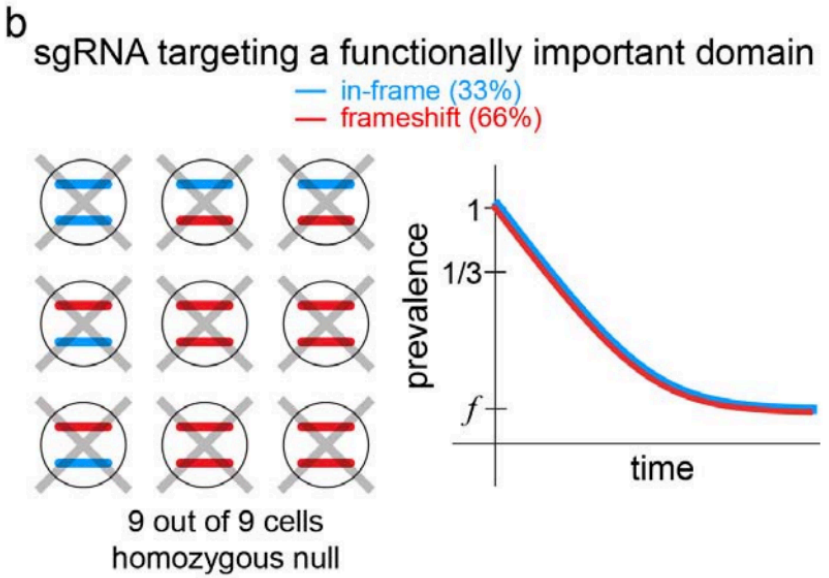
Domain-focused CRISPR screens
Shi et al., 2015, Nat Biotech (Vakoc Lab)





Domain-focused CRISPR screens
Shi et al., 2015, Nat Biotech (Vakoc Lab)

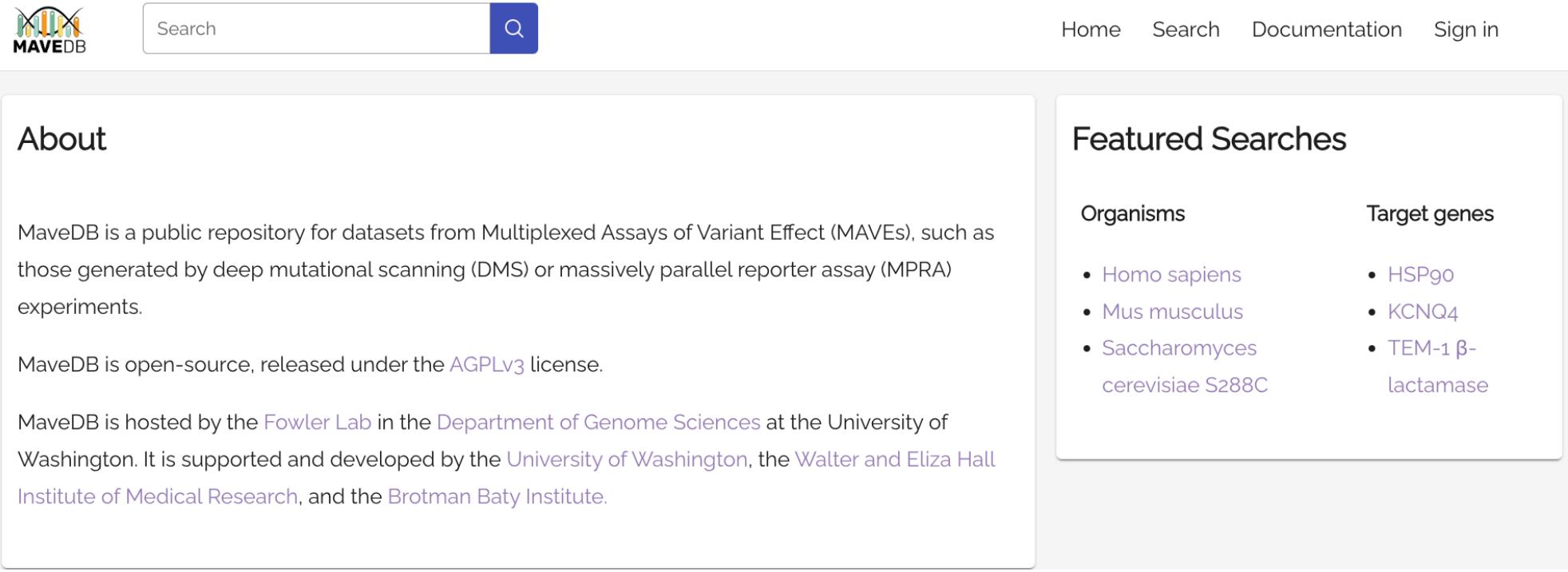




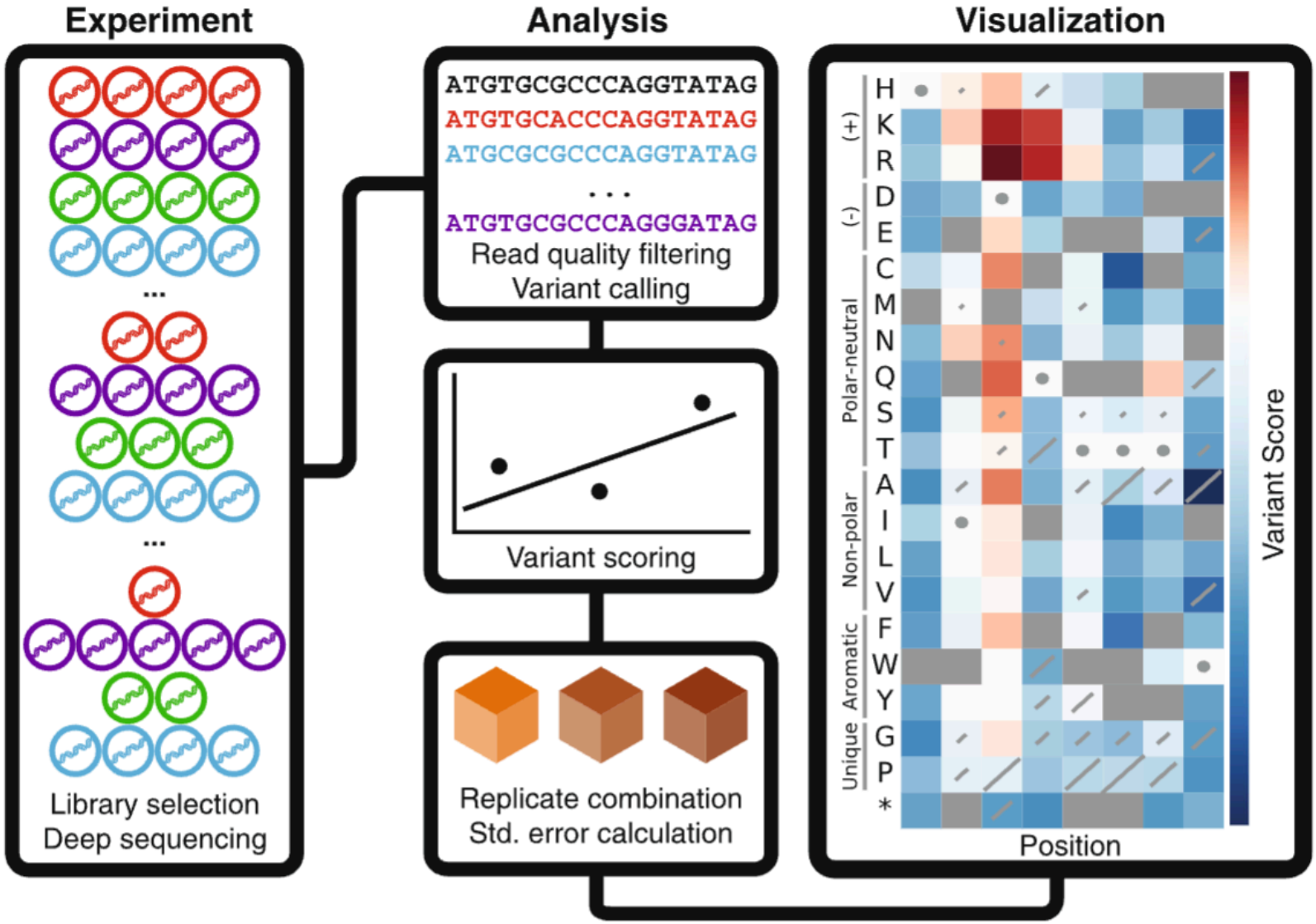
Computational resources for MAVEs (took 60 min to get here)

MaveDB lists data from a rapidly expanding set of DMS, MPRA, and other MAVEs Esposito et al., 2019, Genome Biol (Fowler Lab)

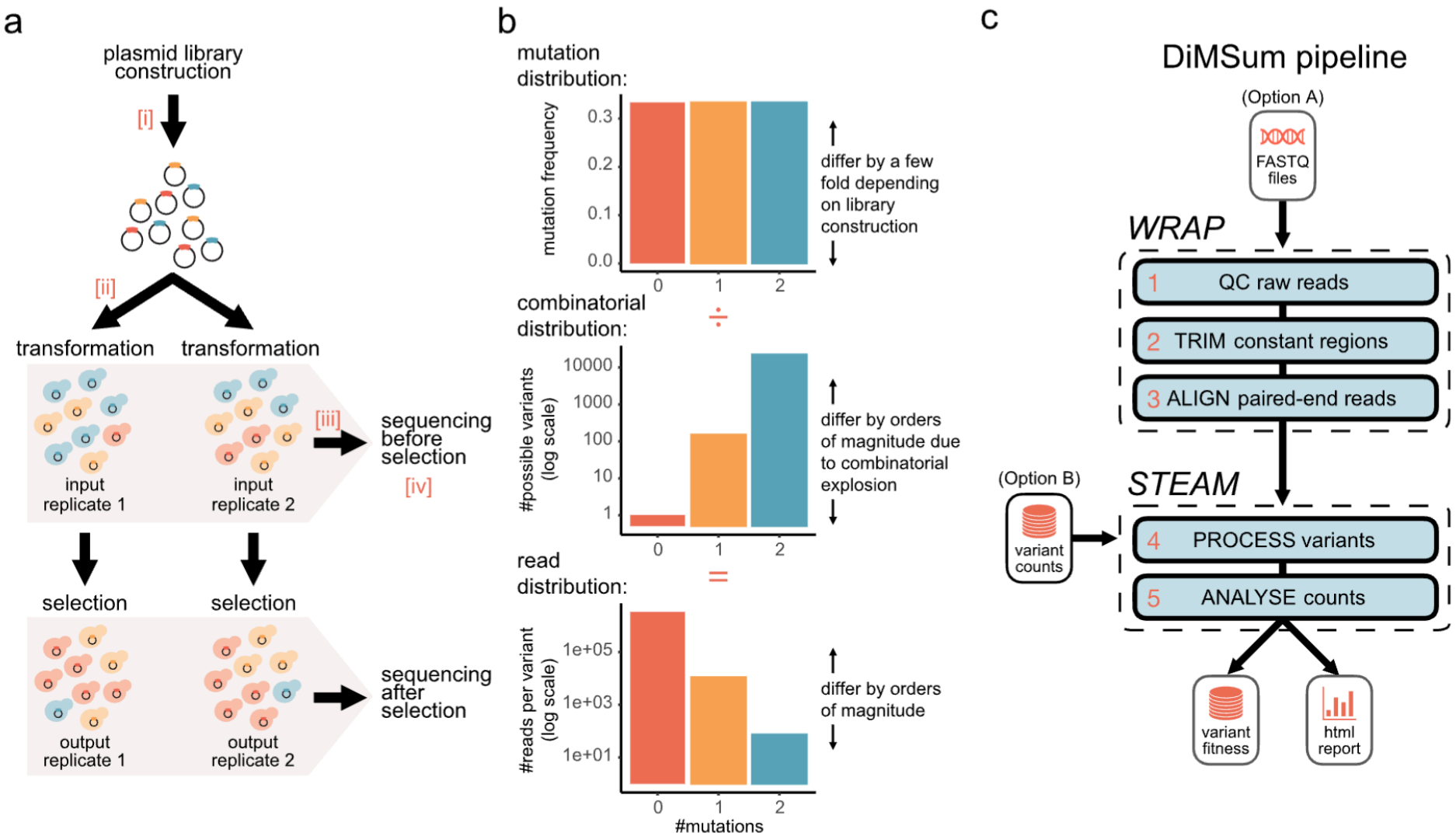
www.mavedb.org



Enrich2: Software for scoring variants in DMS data
Ruben et al., 2017, Genome Biol (Fowler Lab)



DiMSum: Software for scoring variants in DMS data
Faure et al., 2020, Genome Biol (Lehner Lab)



MAVE-NN: Genotype-phenotype maps from multiplex assays of variant effect

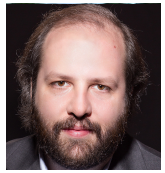
---

Tareen A, Posfai A, Ireland WT, McCandlish DM, JBK
Genome Biol (2022)


Ammar Tareen

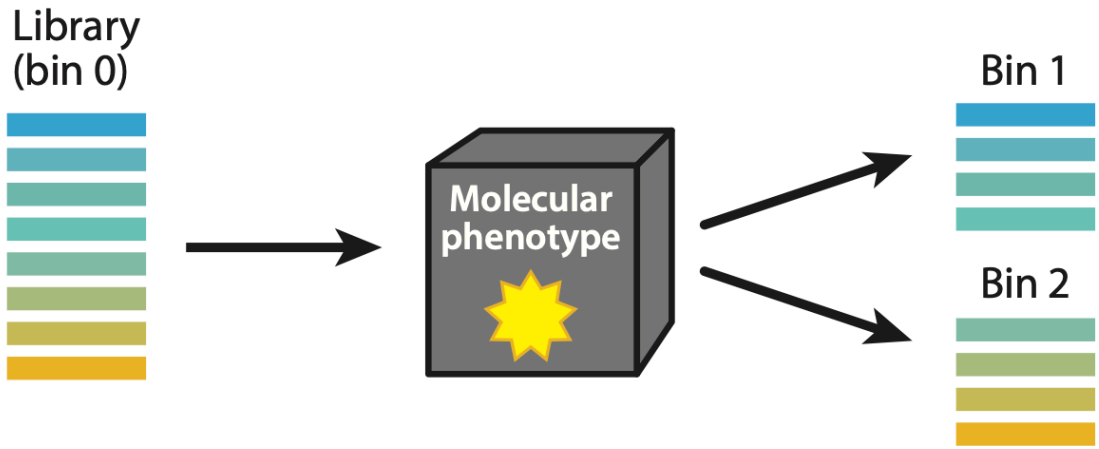

Anna Posfai


Bill Ireland
(Caltech / Harvard)


David McCandlish

A general computational framework for quantitatively modeling MAVE data

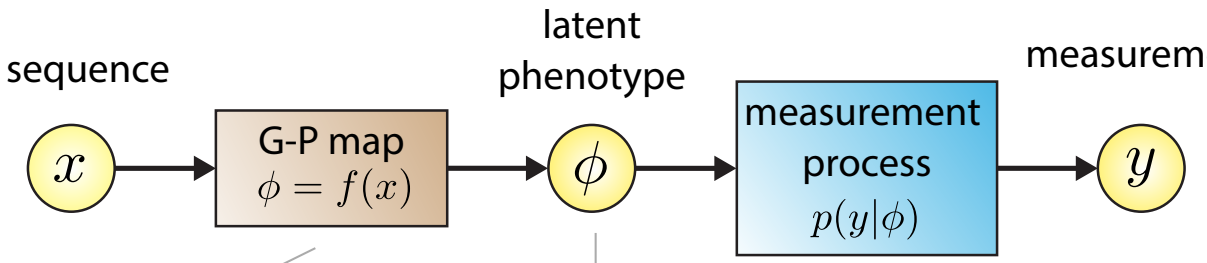
MAVES have a common form that is useful to consider when quantitatively modeling their data.



Kinney and McCandlish, 2019, *Ann Rev G Hum Genet*

MAVE-NN implements a unified approach for modeling arbitrary genotype-phenotype maps from all types of MAVE experiments.

latent phenotype model



genotype-phenotype (G-P) map

A deterministic function that compresses the relevant information in each sequence into a low-dimensional latent phenotype.

latent phenotype

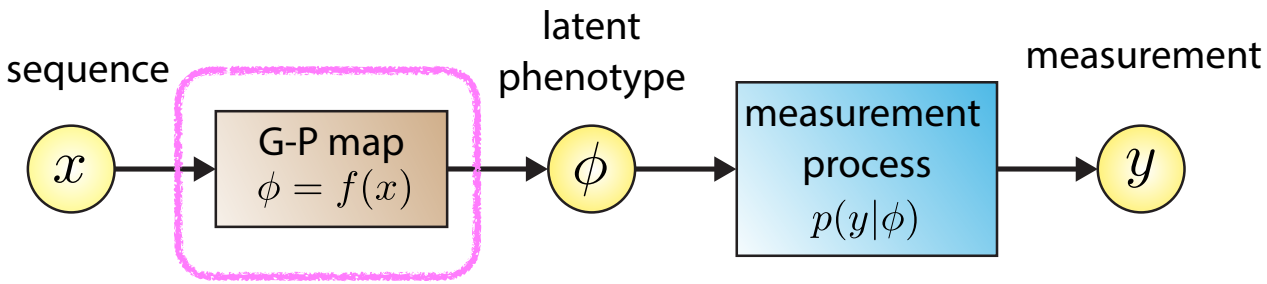
A quantity of biological importance that influences measurements but is not directly observed.

measurement process

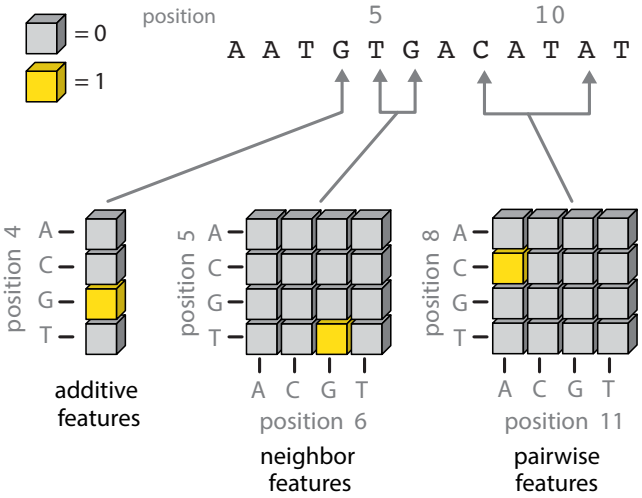
A stochastic mapping (typically noisy and nonlinear) from the latent phenotype to experimental measurements.

Tareen et al., 2022, *Genome Biol*

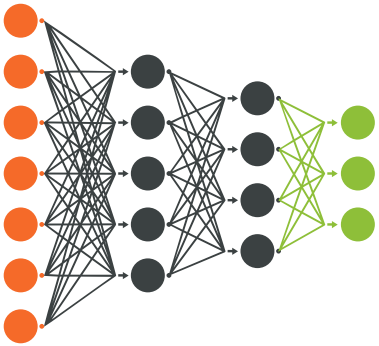
MAVE-NN supports diverse types of models for genotype-phenotype maps.



linear feature-based models

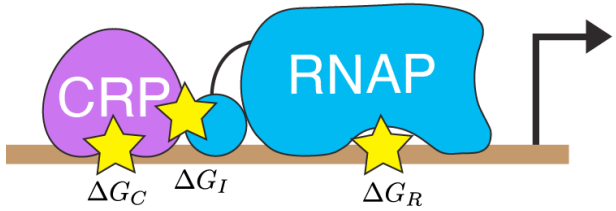


black box neural networks



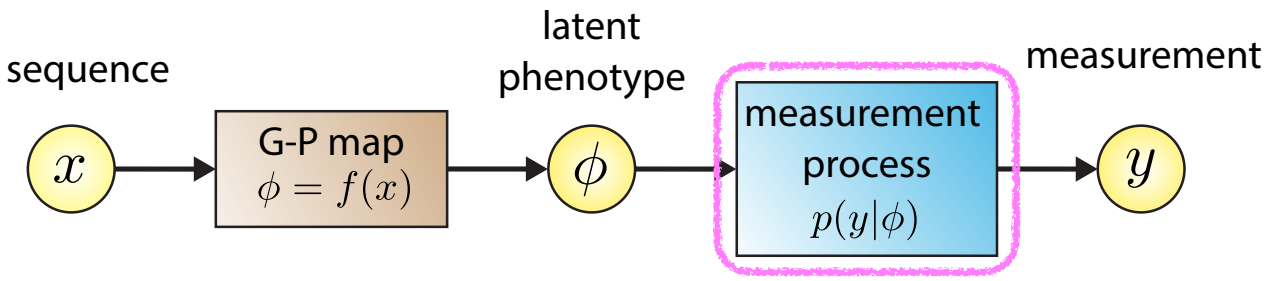
custom genotype-phenotype maps

(e.g. biophysical models)

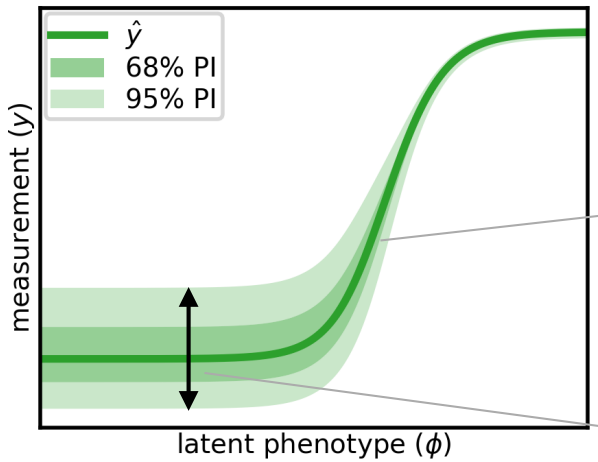


Tareen et al., 2022, *Genome Biol*

MAVE-NN supports both continuous and discrete experimental measurements.



global epistasis (GE) measurement process

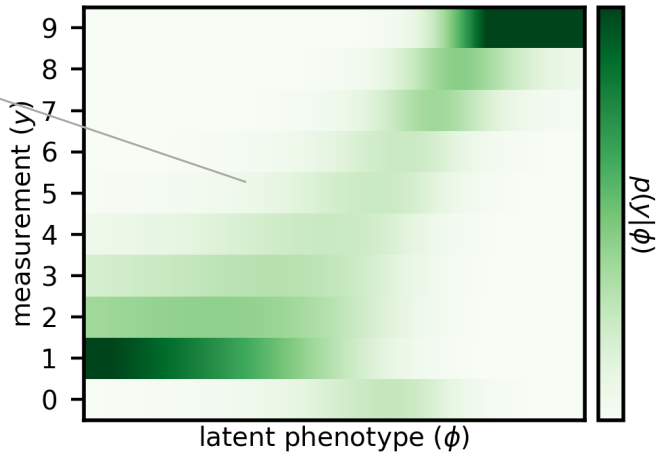


conditional probability
flexible neural network

nonlinearity
sum of sigmoids

noise model
heteroscedastic
Skewed-t
(Jones & Faddy, 2003)

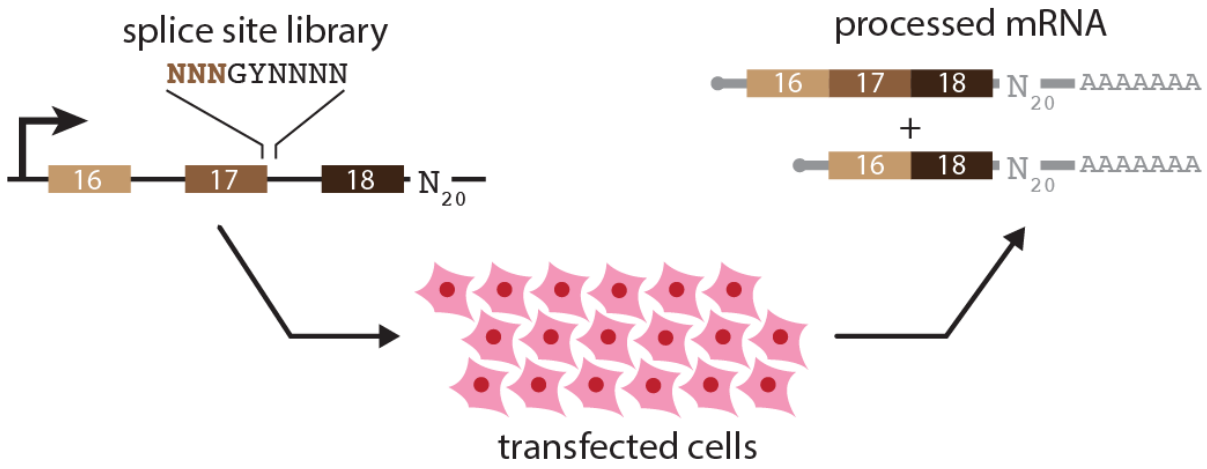
measurement process agnostic (MPA) measurement process



Tareen et al., 2022, *Genome Biol*

Example: RNA-seq MPRA of 5’ splice sites in human cells.

massively parallel splicing assay (MPSA) in HeLa cells
(exon inclusion for ~33K 5’ splice sites)



splice site sequence	percent spliced in (PSI)
CCGGUUUGC	0.4%
ACGGUCUGA	0.5%
AUGGUAAGA	99.6%
CCGGCACGG	0.5%
CGGGCAAGG	0.9%
TCGGUAAGU	139.7%
ACGGUAAGA	111.4%

Wong et al, 2018, *Mol Cell*

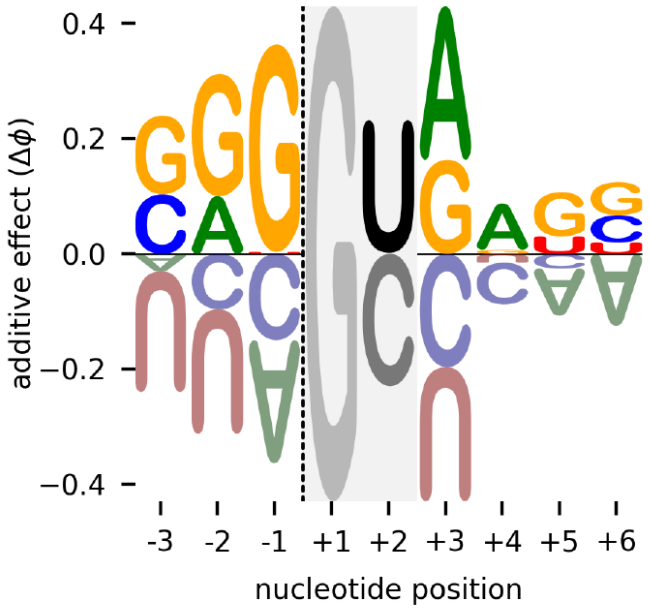
Example: RNA-seq MPRA of 5’ splice sites in human cells.

pairwise genotype-phenotype map

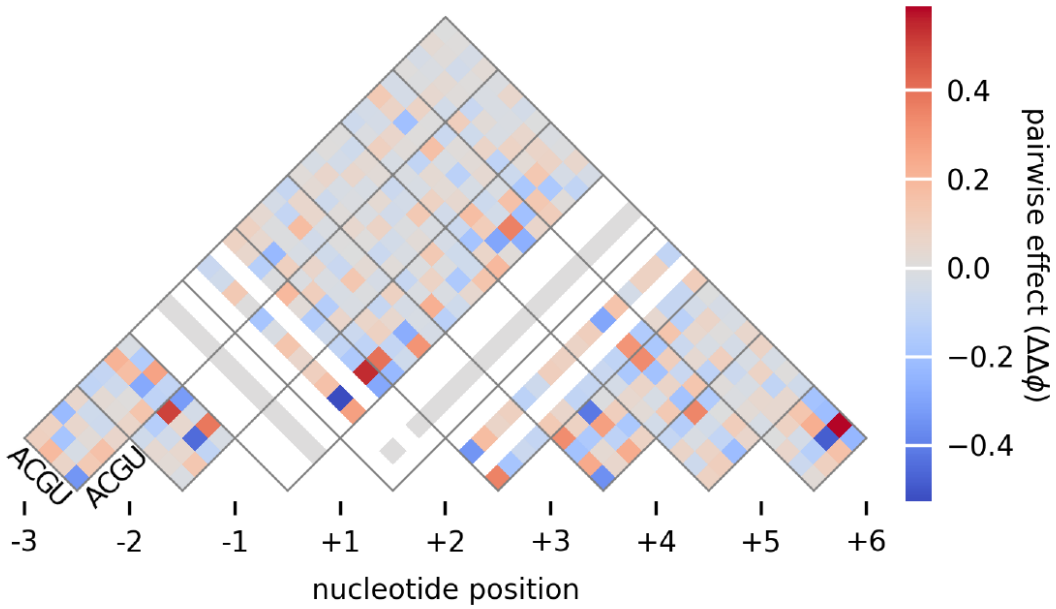
$$\phi(x) = \theta_0 + \sum_{l,c} \theta_{l:c} x_{l:c} + \sum_{l,l',c,c'} \theta_{l:c,l':c'} x_{l:c} x_{l':c'}$$

$x_{l:c}$: one-hot encoding of sequence

$\theta_{l:c}$: additive parameters



$\theta_{l:c,l':c'}$: pairwise parameters



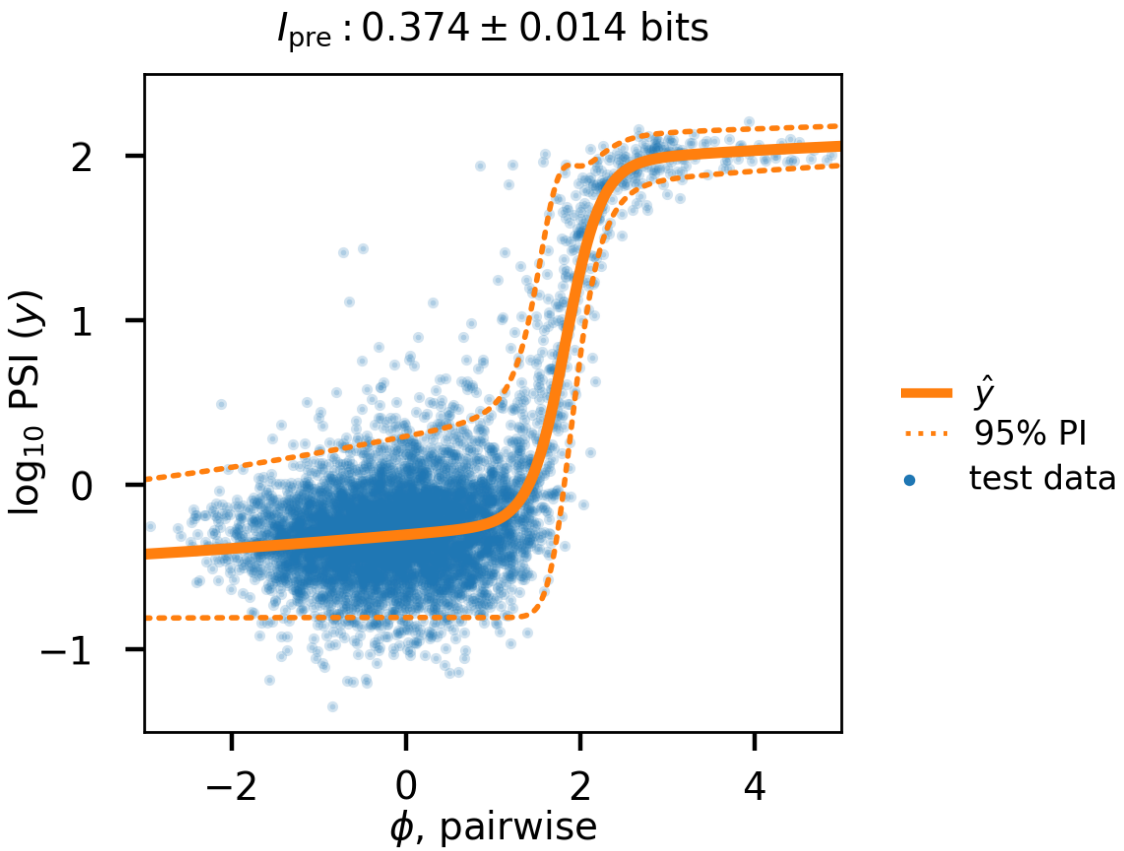
Tareen et al., 2022, *Genome Biol*

Example: RNA-seq MPRA of 5’ splice sites in human cells.

GE measurement process

mode shape parameters

$$p(y \mid \phi) = \text{SkewedT}(y \mid \hat{y}(\phi),\ \eta(\hat{y}))$$

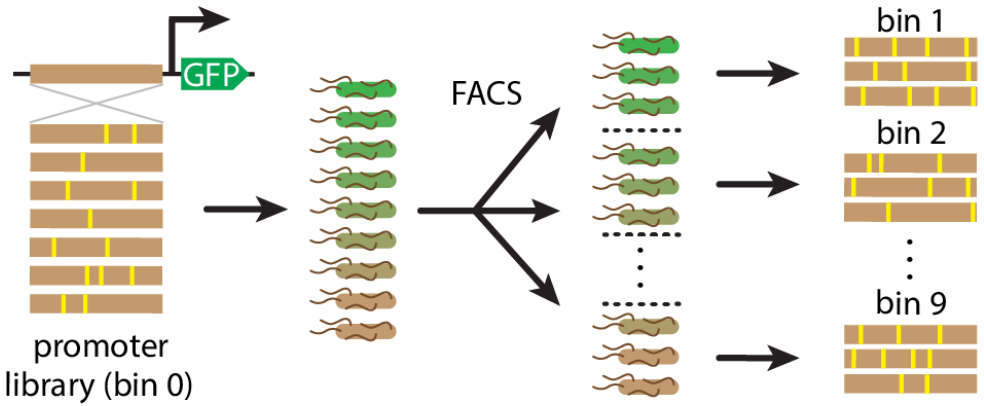


Tareen et al., 2022, *Genome Biol*

Example: Sort-seq MPRA in bacterial promoters.

mutagenic study of the *Escherichia coli lac* promoter

(~50K variants sorted into 10 bins)

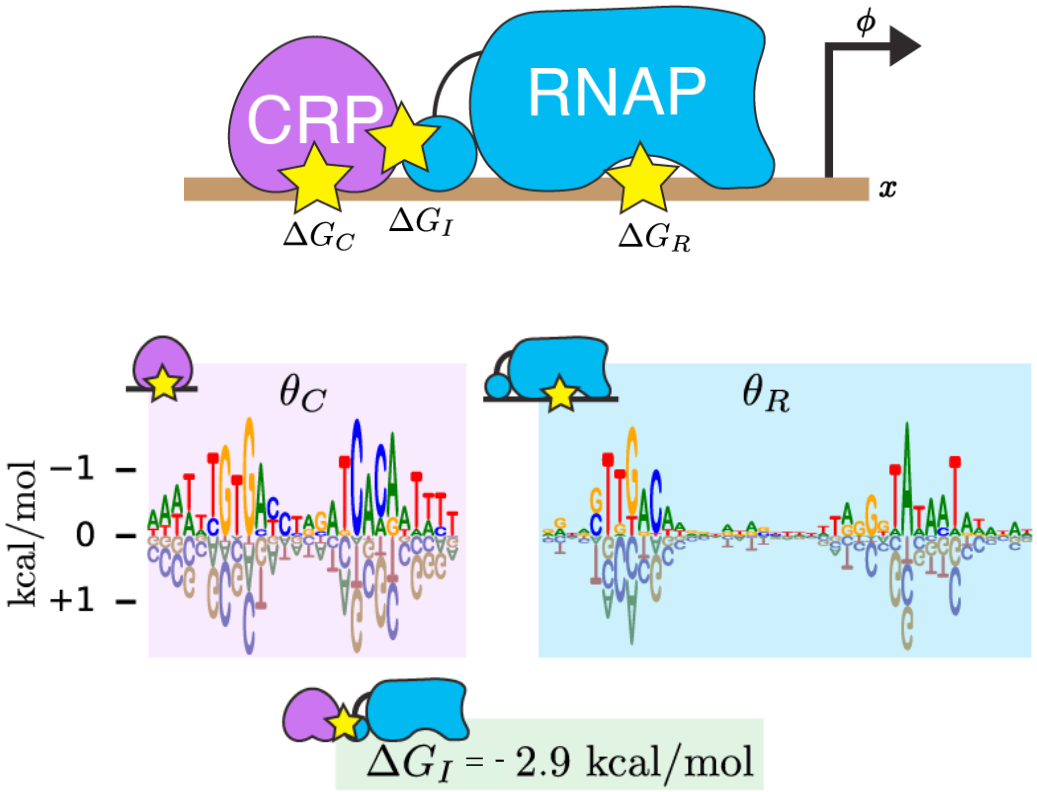


promoter sequence	read count			
	bin 0	bin 1	...	bin 9
	0	1		0
	1	0		0
	0	0		2
	1	0		1
	1	0		0
	0	3		1
	2	0		0

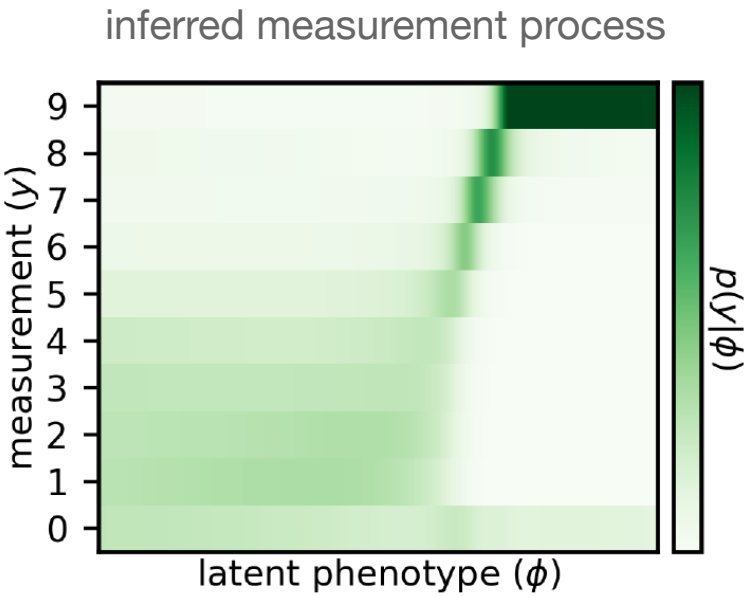
Kinney et al., 2010, *PNAS*

Example: Sort-seq MPRA in bacterial promoters.

thermodynamic genotype-phenotype map



MPA measurement process



Kinney 2010: ~1 week on 50 nodes

MAVE-NN: ~15 min on a laptop

Tareen et al., 2022, *Genome Biol*

PNAS

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY
APPLIED MATHEMATICS

OPEN ACCESS

Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power

Peter D. Tonner[a,1], Abe Pressman[b], and David Ross[b]

Edited by Wing Hung Wong, Stanford University, Stanford, CA; received August 3, 2021; accepted March 4, 2022

Faure and Lehner *Genome Biology* (2024) 25:303
https://doi.org/10.1186/s13059-024-03444-y

Genome Biology

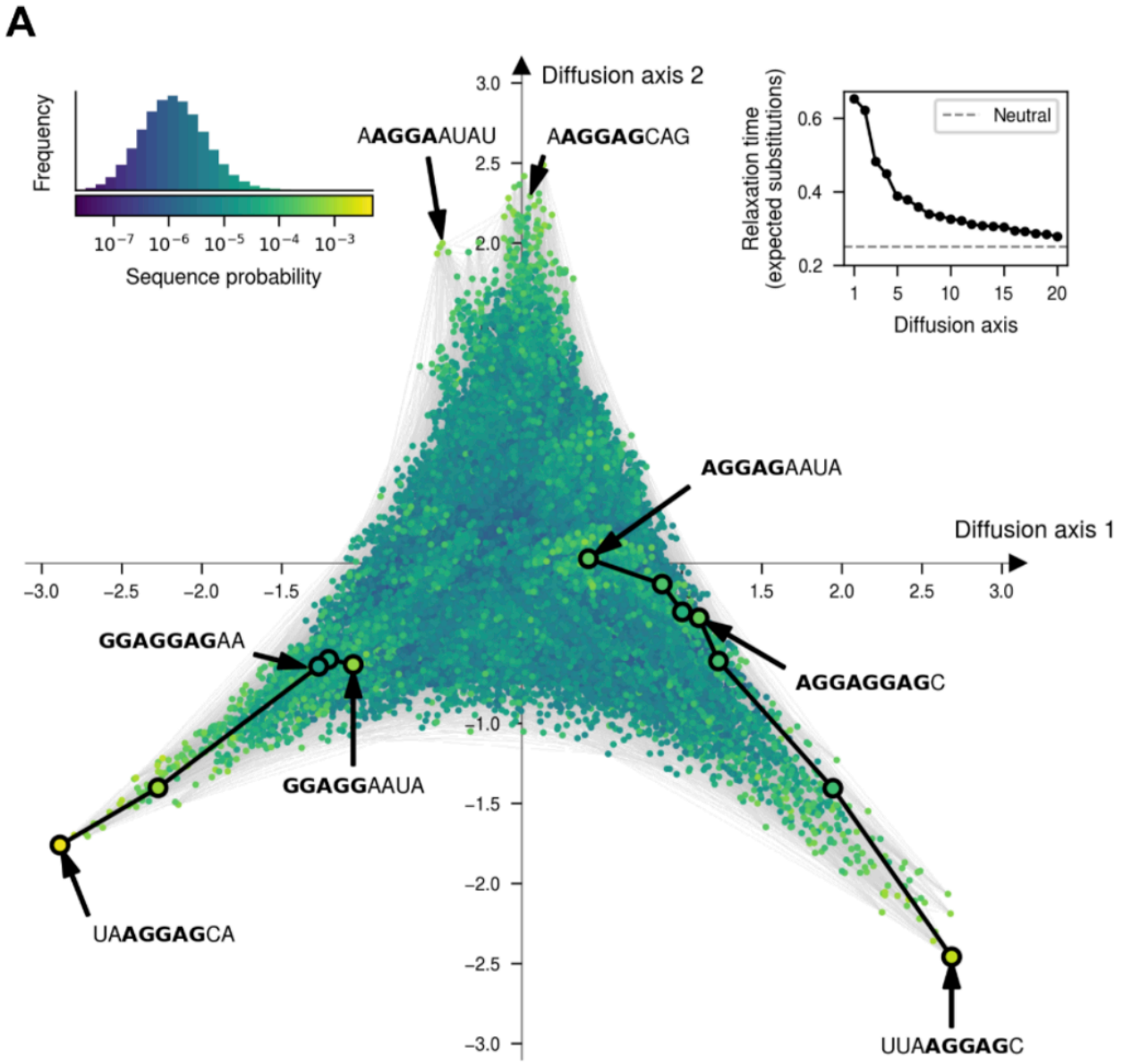
METHOD

Open Access

MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data

Andre J. Faure[1,5*] and Ben Lehner[1,2,3,4*]

gpmap-tools
Martí-Gómez et al., 2025, bioRxiv (McCandlish Lab)



Kinney Lab

Lab members

Andalus Ayaz
Aiden Cordero
Jack Desmarias
Zhihan Liu
Yuma Ishigami
Debora Tenenbaum

Lab alumni

Evan Seitz
Mahdi Kooshkbaghi
Anna Posfai
Mandy Wong
Ammar Tareen
Wei-Chia Chen

Rhys Adams
Talitha Forcier
Manraj Gill
Kush Koshic
Daniel Jones

Honorary lab members

(Krainer Lab, CSHL)
Carlos Martí
(McCandlish Lab, CSHL)
Rebecca Rousseau
(Phillips Lab, Caltech)

Collaborators

Adrian Krainer (CSHL)
David McCandlish (CSHL)
Peter Koo (CSHL)
Bryce Nickels (Rutgers)
Rob Phillips (Caltech)

Funding

NIH R35GM133777
NIH R01HG011787
Northwell Health
Moore Foundation