

Massively parallel assays



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

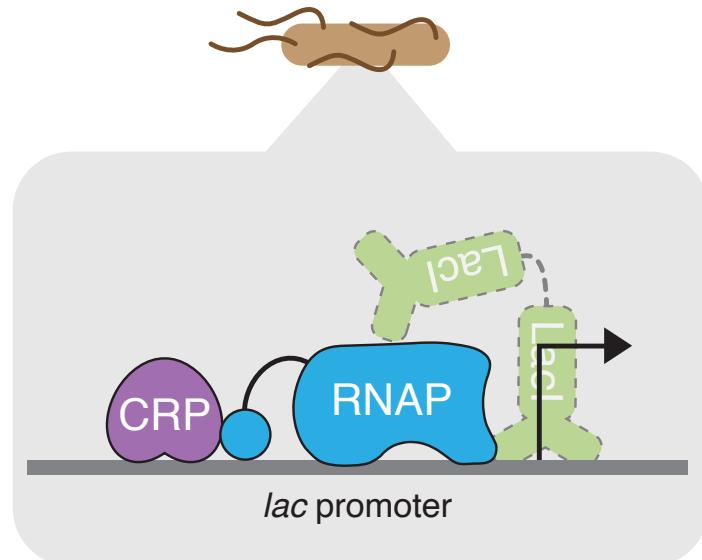
Advanced Sequencing Technologies & Bioinformatics Analysis
CSHL Course
14 November 2024

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

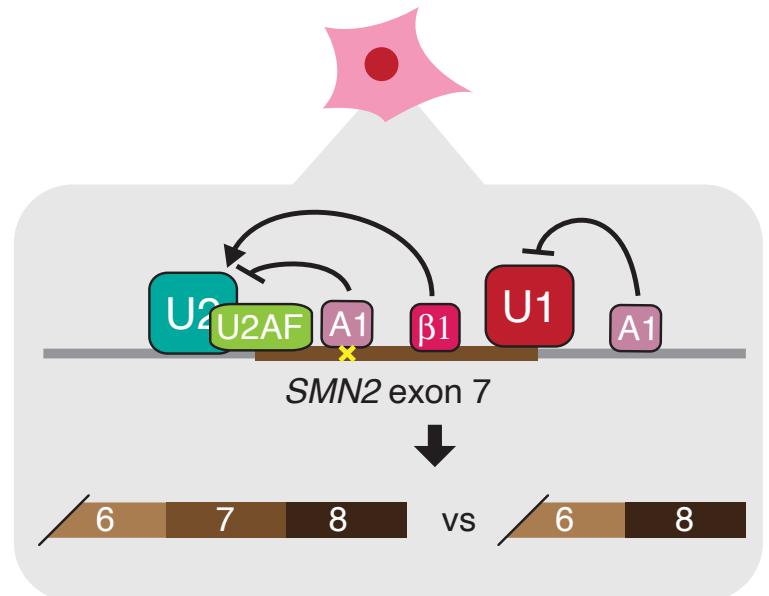
Our work involves a tightly knit combination of experiment, computation, and mathematical theory.

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:

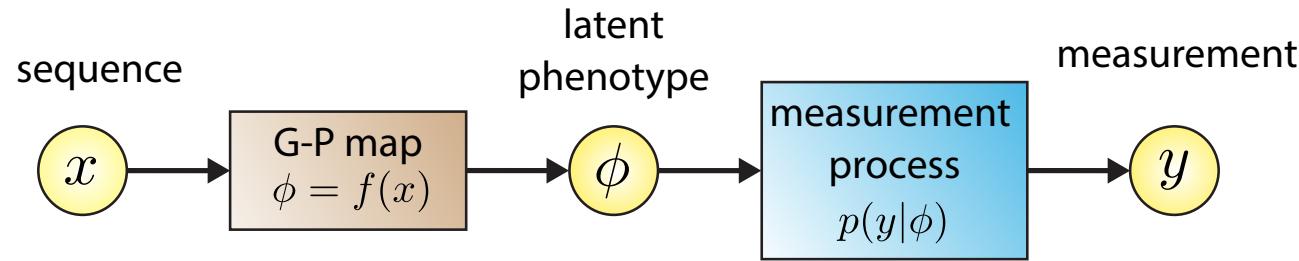
transcriptional regulation in bacteria



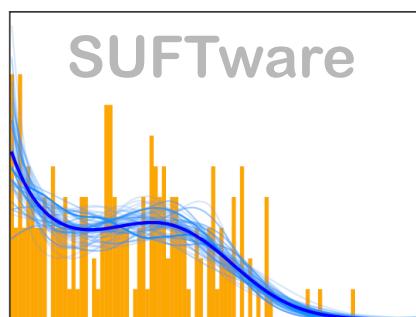
alternative mRNA splicing in human cells



Our **theoretical and computational work** develops methods for analyzing the data produced by MPRA^s 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.



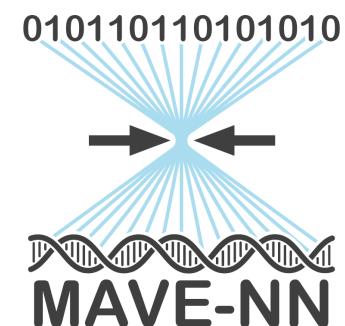
These efforts include devising new mathematical and computational approaches, as well as developing robust and efficient software for use by the larger genomics community.



<https://software.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

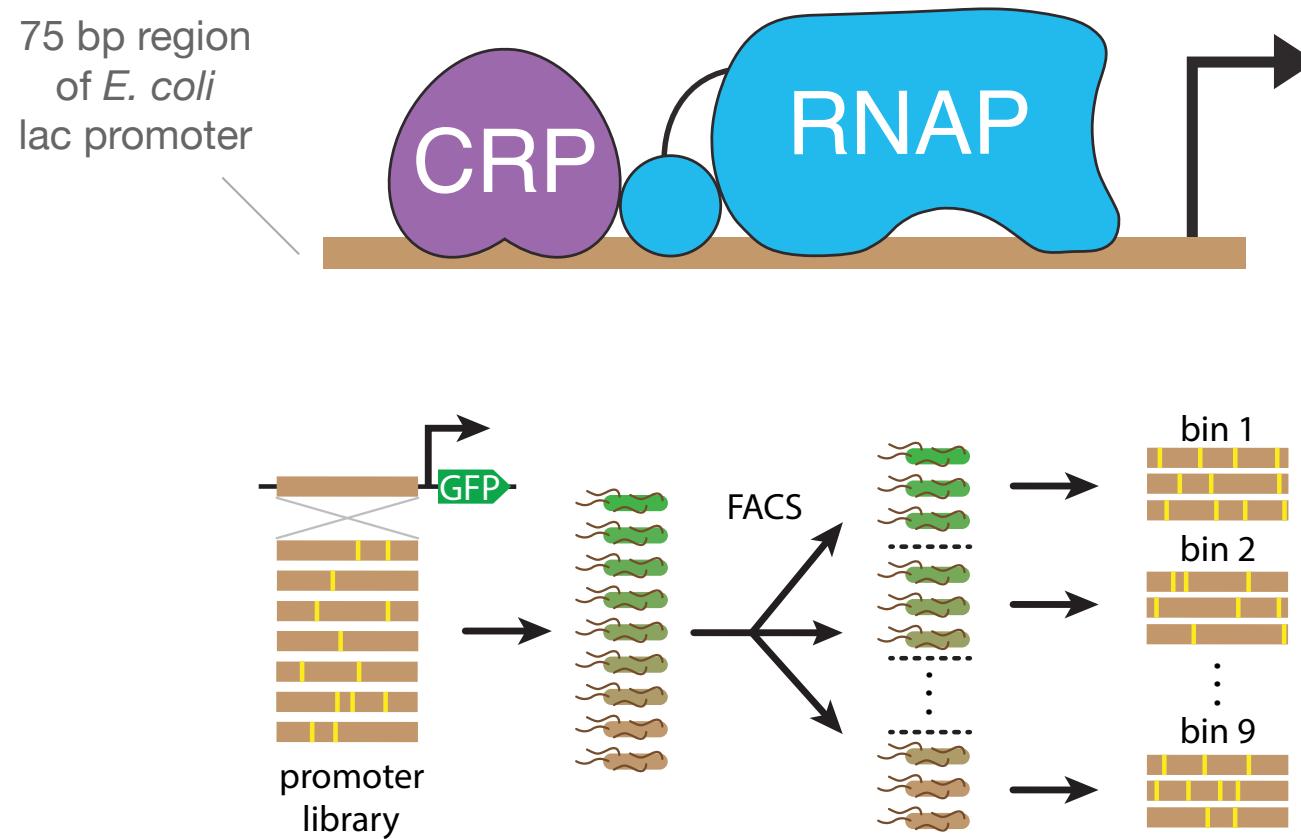


Curtis Callan
(Princeton)



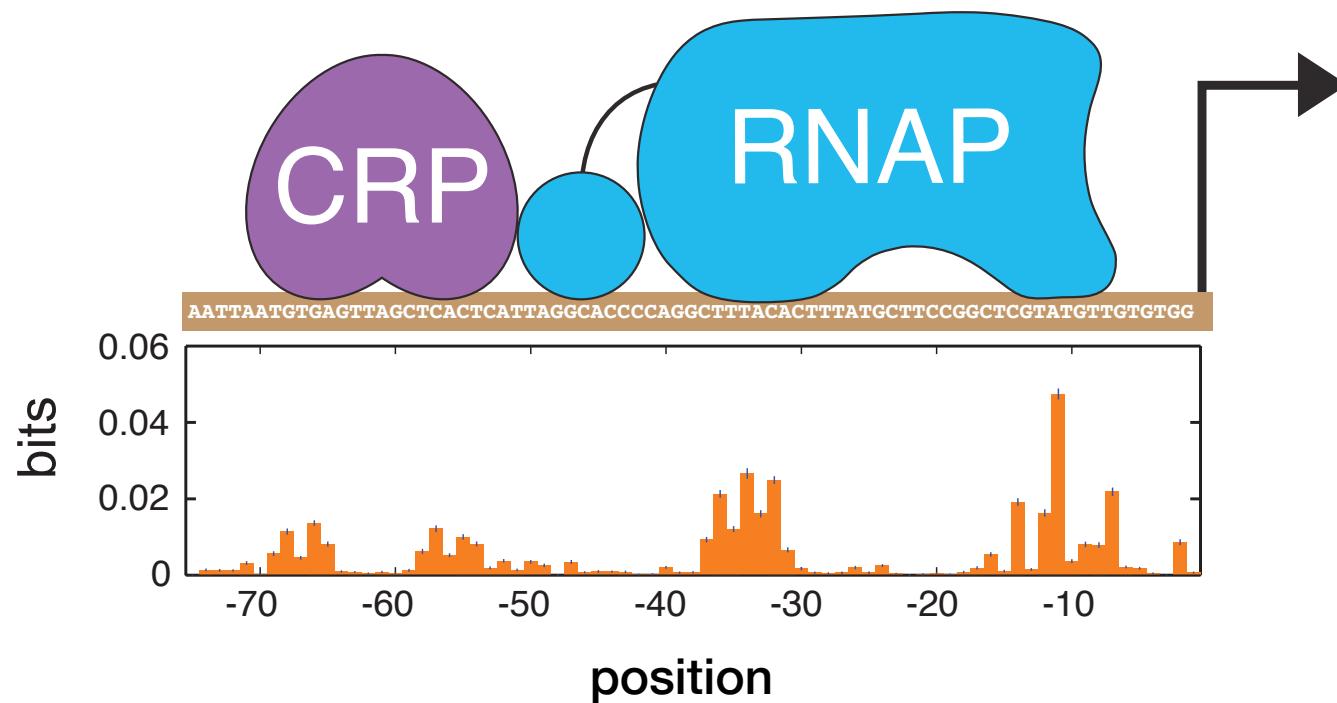
Edward Cox
(Princeton)

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



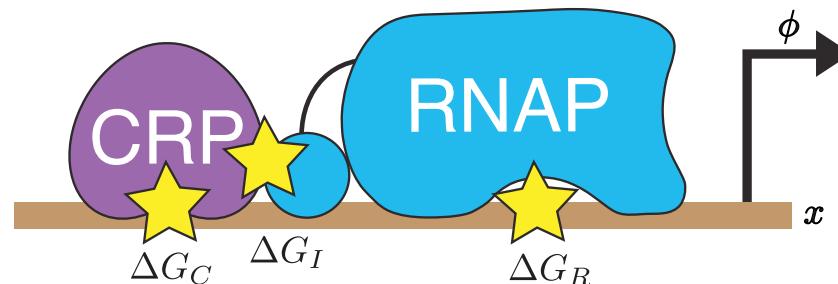
Simultaneous measurements for ~250,000 different promoters

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

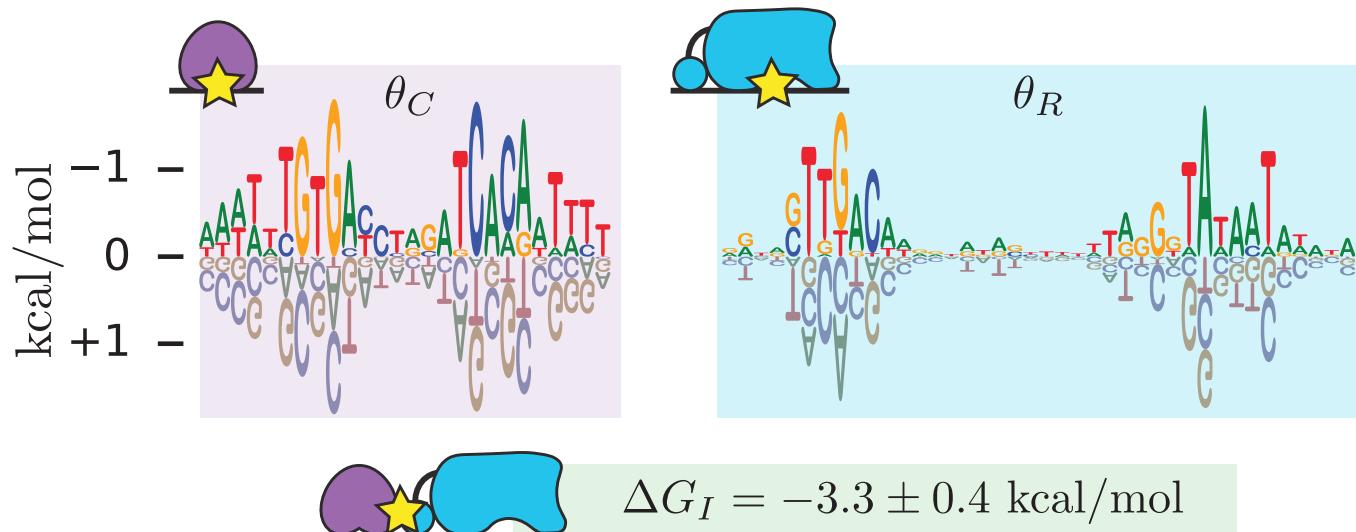


Fitting the parameters of biophysical models to Sort-seq data allows one to quantitatively define the biophysical mechanisms of gene regulation.

proposed model:



inferred parameters:



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

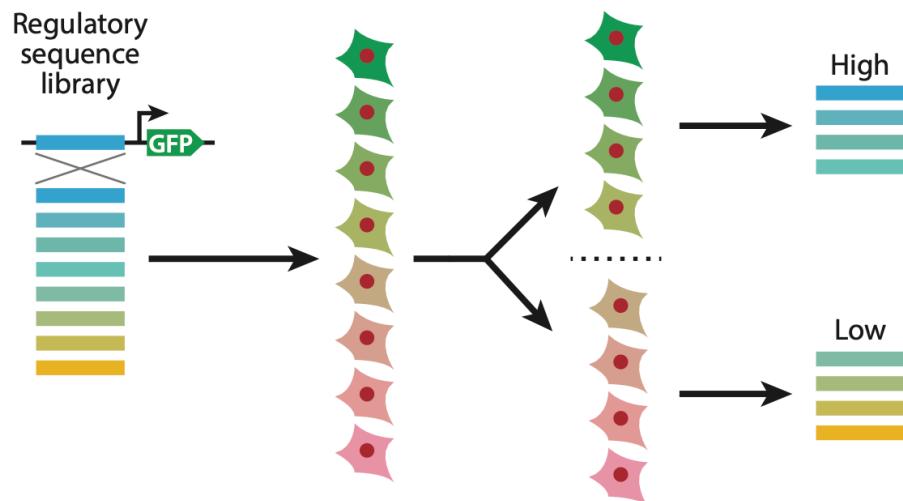


Annual Review of Genomics and Human Genetics
Massively Parallel Assays
and Quantitative
Sequence–Function
Relationships

Justin B. Kinney and David M. McCandlish

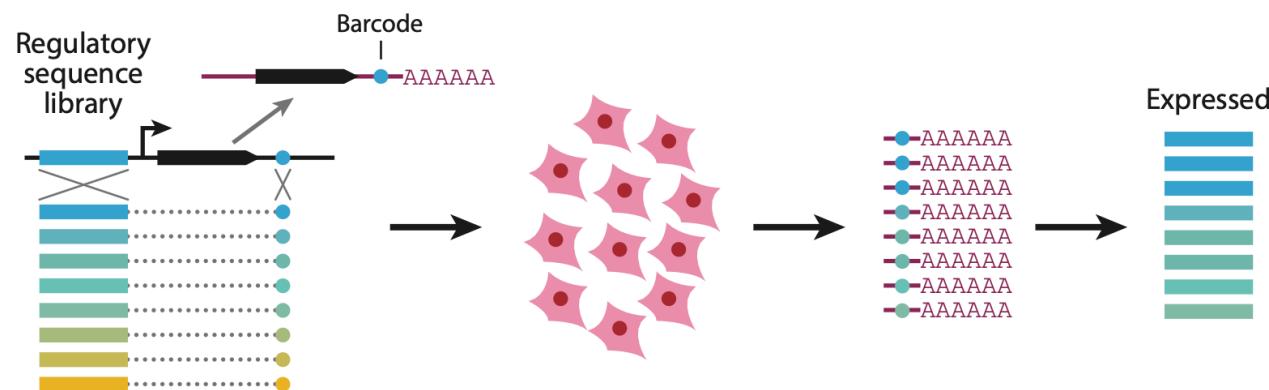
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; email: jkinney@cshl.edu, mccandlish@cshl.edu

d Sort-seq MPRA



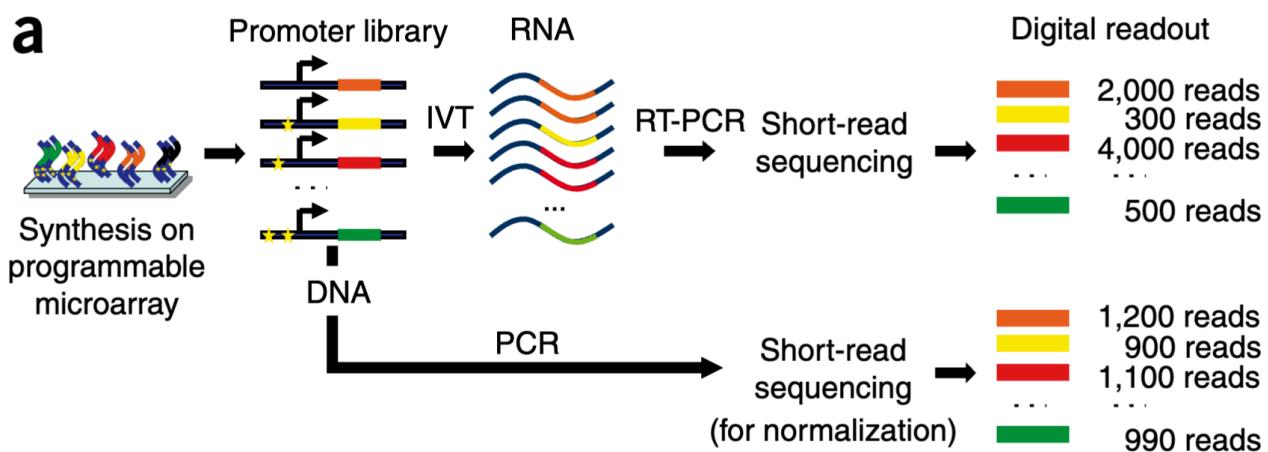
Massively parallel reporter assays (MPRAs)

e RNA-seq MPRA

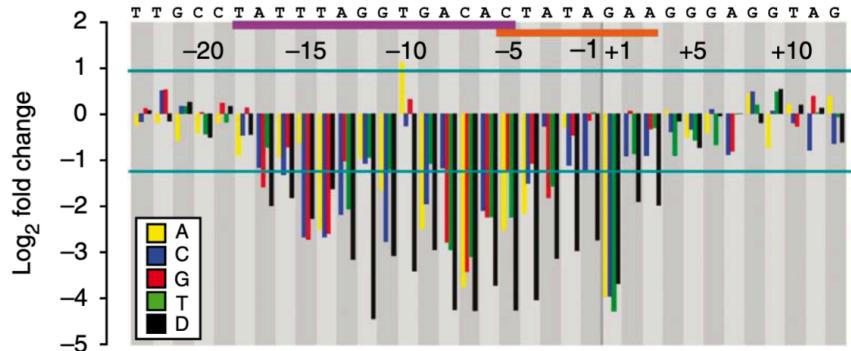
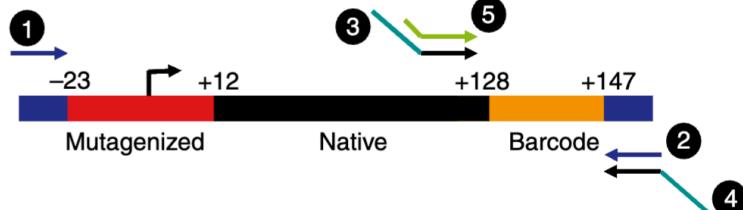


MPRAs on promoters in vitro

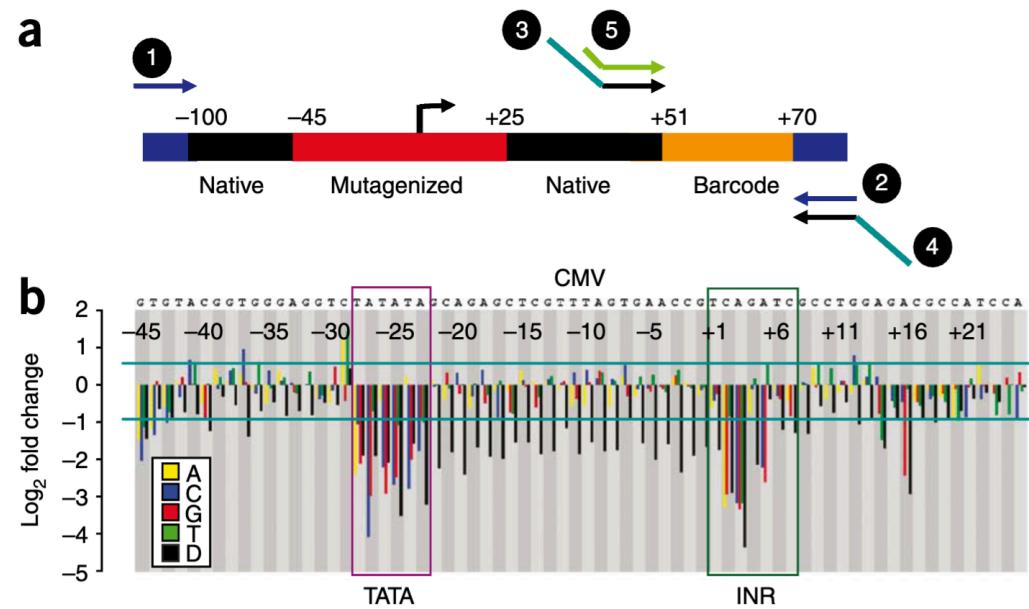
Patwardhan et al., 2009, Nat Biotech (Shendure Lab)



bacteriophage RNA polymerase

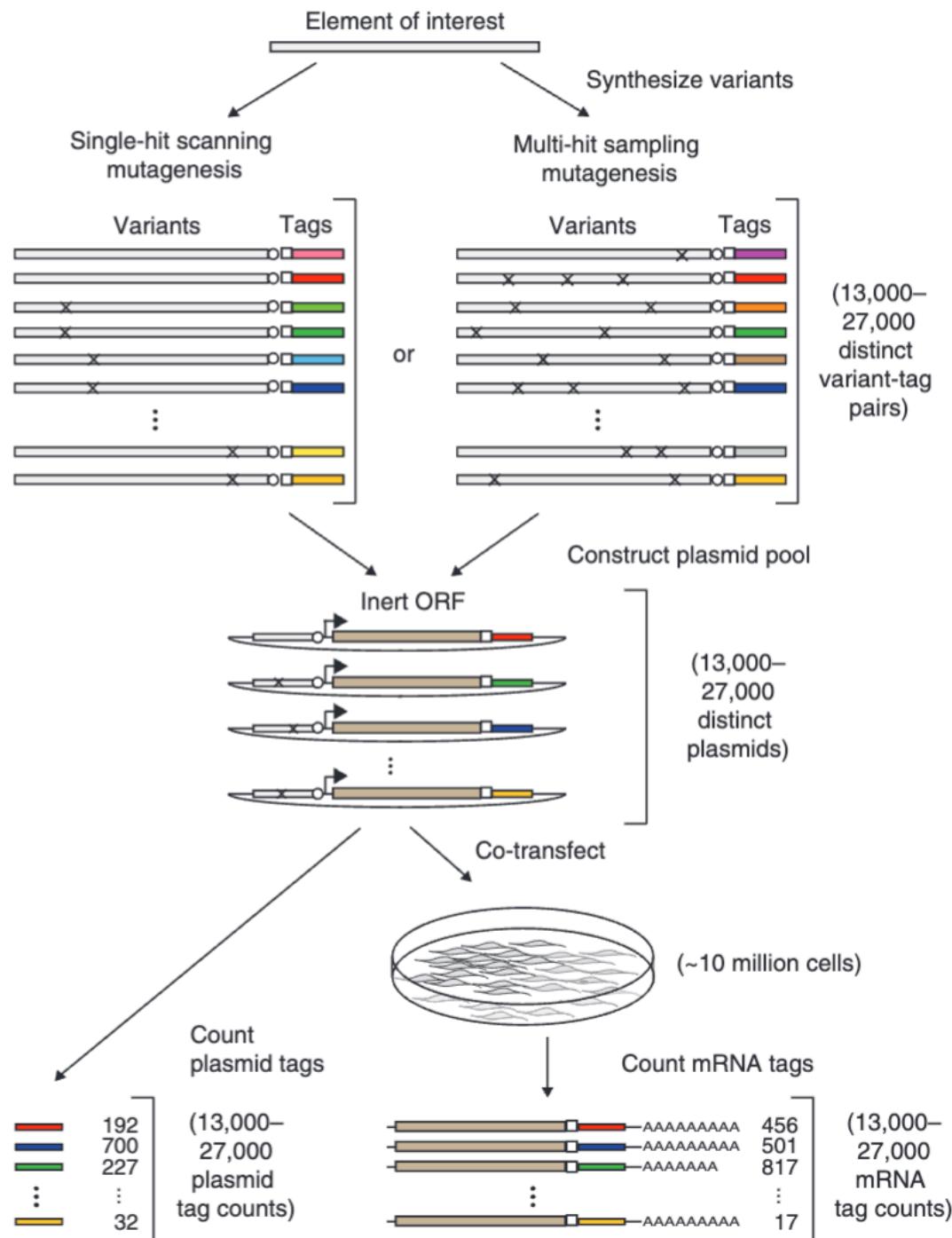


HeLa nuclear extracts



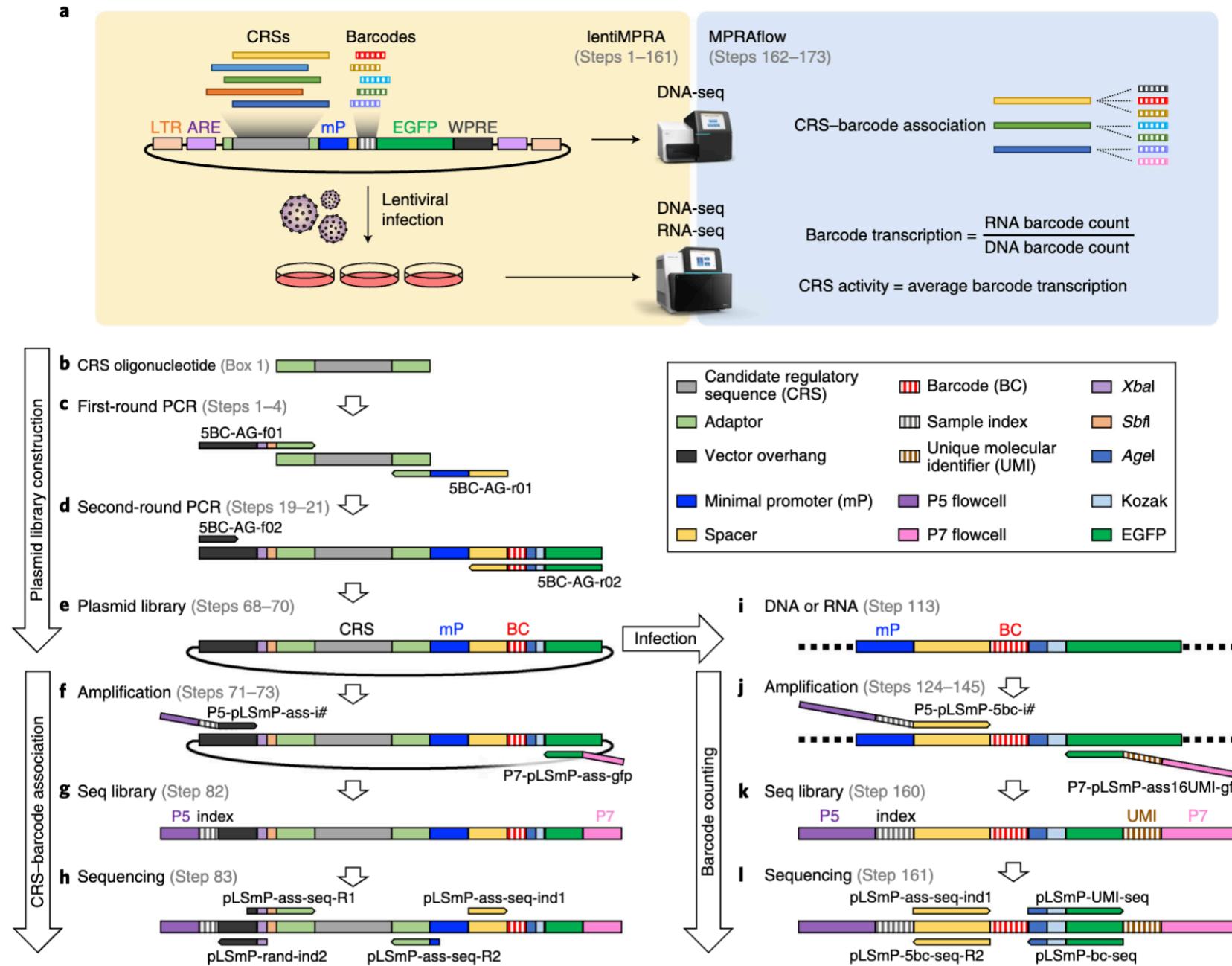
MPRAs on enhancers in human cells

Melnikov et al., 2012, Nat Biotech (Melnikov Lab)



LentiMPRA

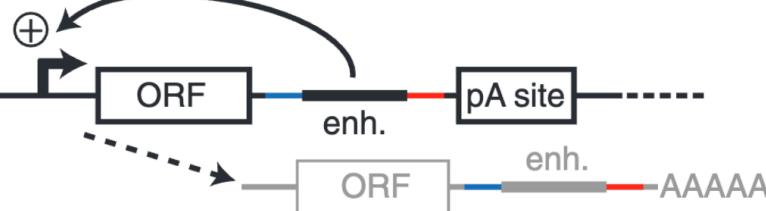
Gordon et al., 2020, Nat Protocol (Ahituv Lab)



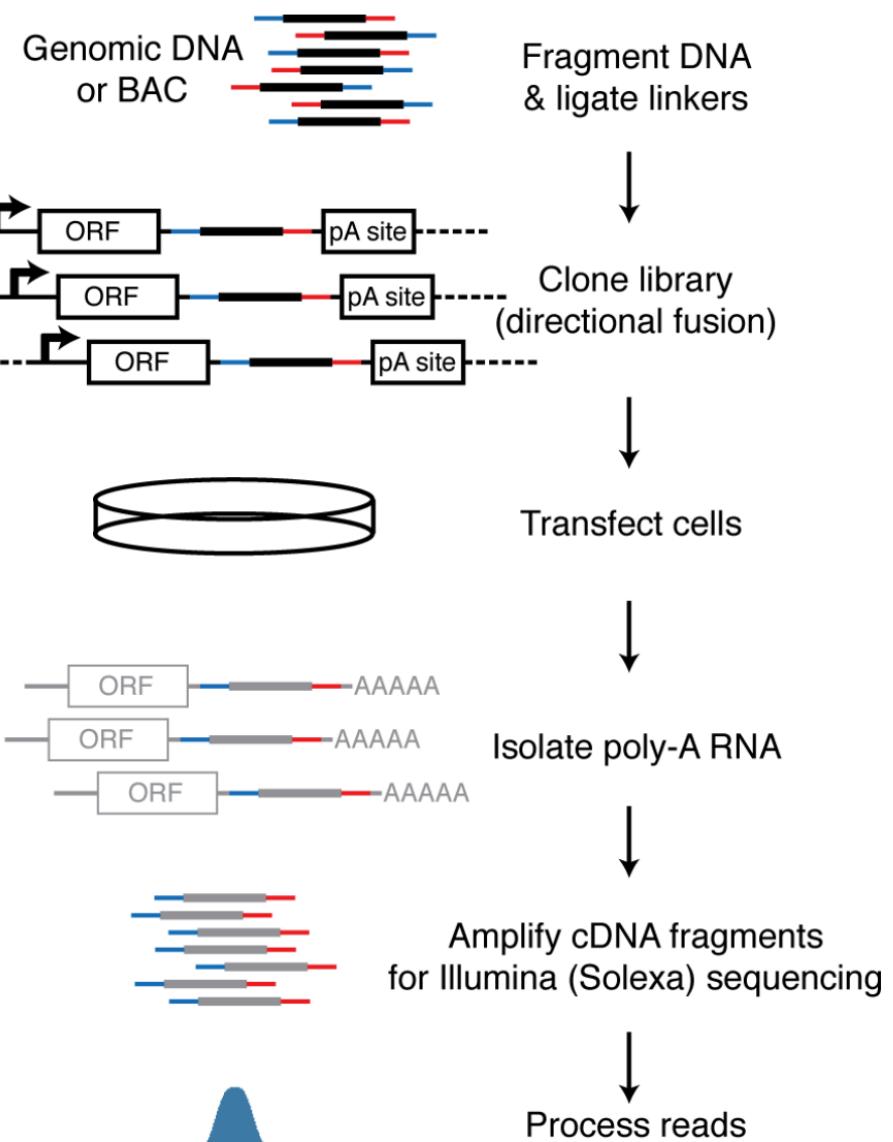
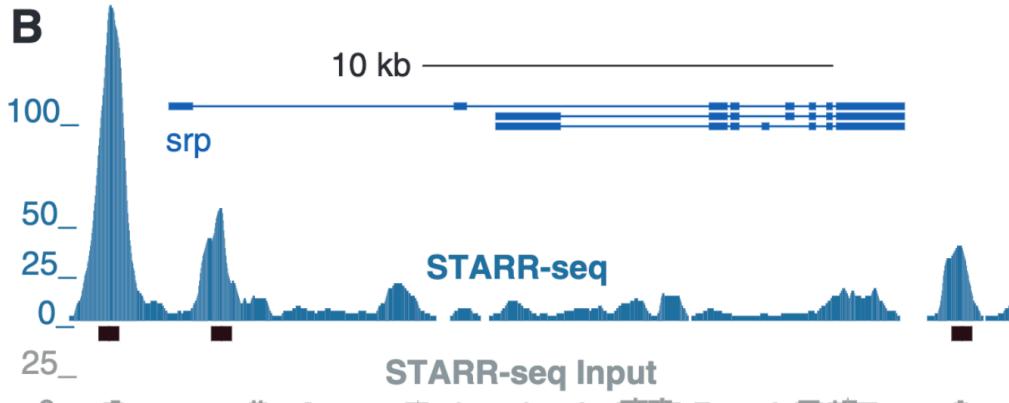
STARR-Seq

Arnold et al., 2014, Science (Stark Lab)

A



B



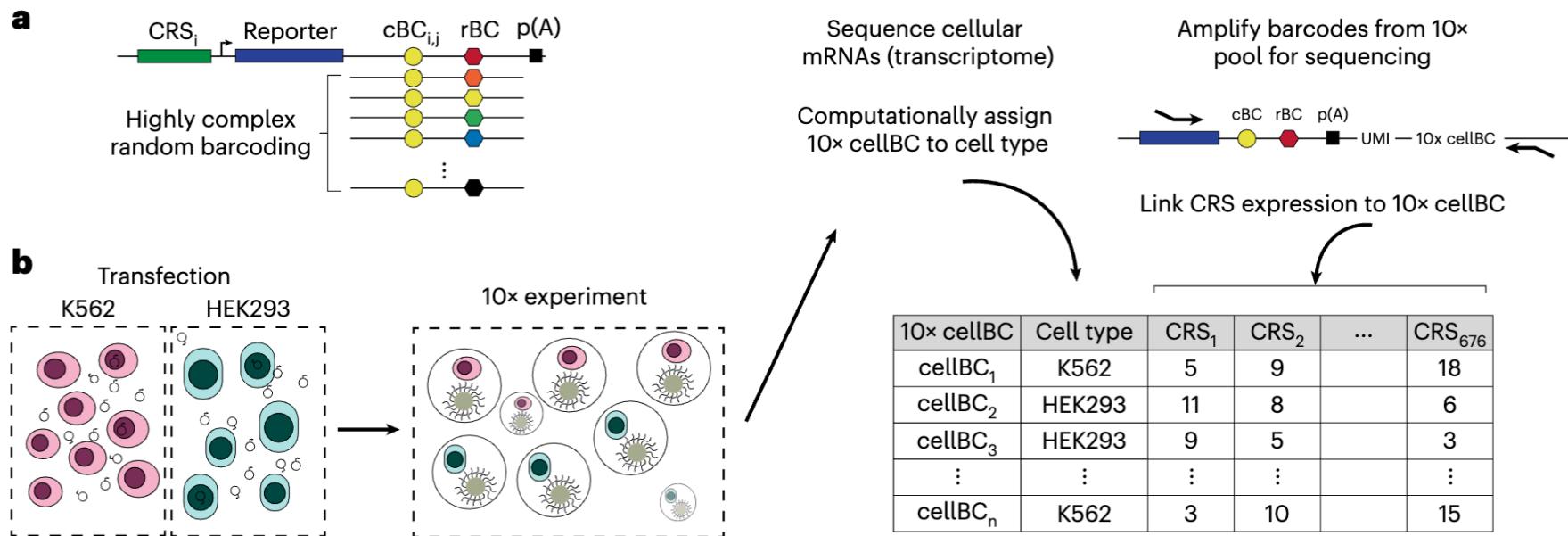
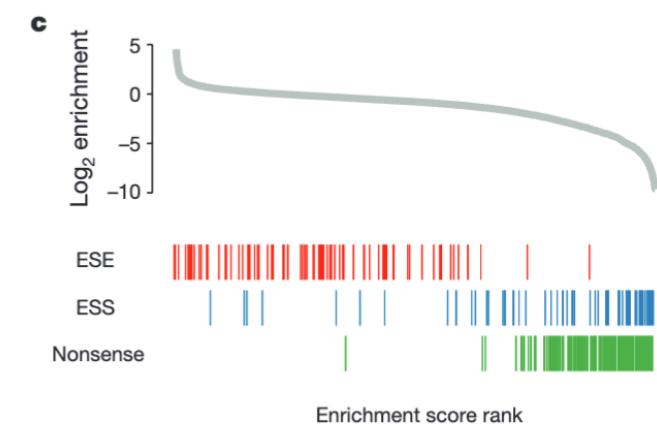
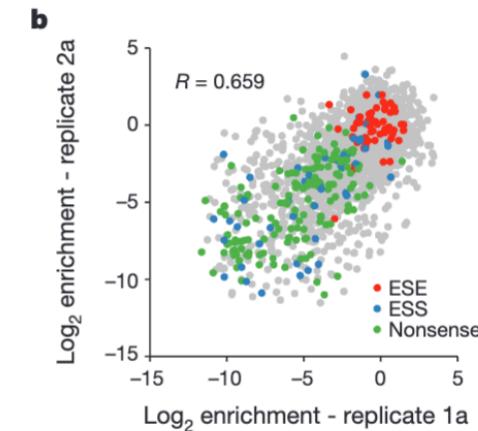
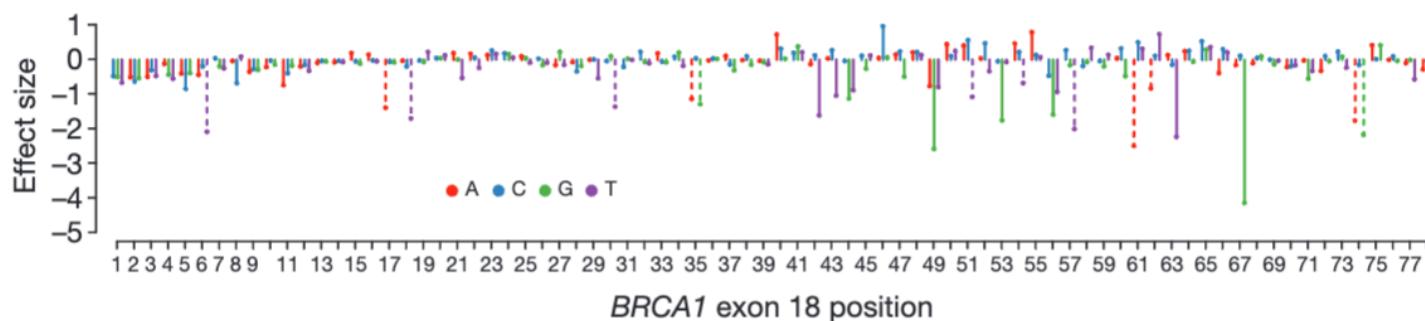
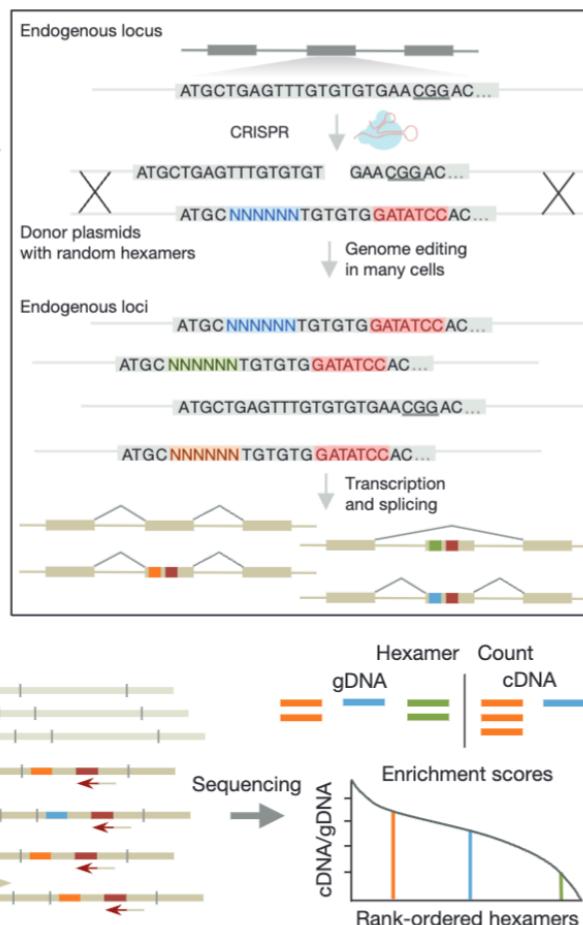
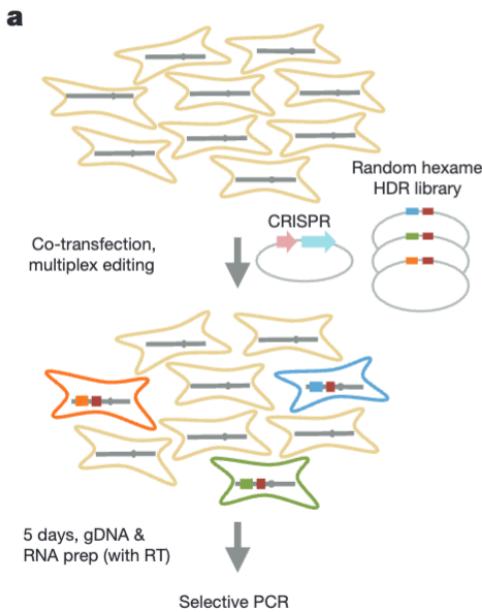


Fig. 1 | scMPRA measures CRS activity at single-cell resolution. **a**, Each CRS reporter construct is barcoded with a cBC that specifies the identity of the CRS, and a highly complex rBC. The complexity of the cBC–rBC pair ensures that the probability of identical plasmids being introduced into the 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.

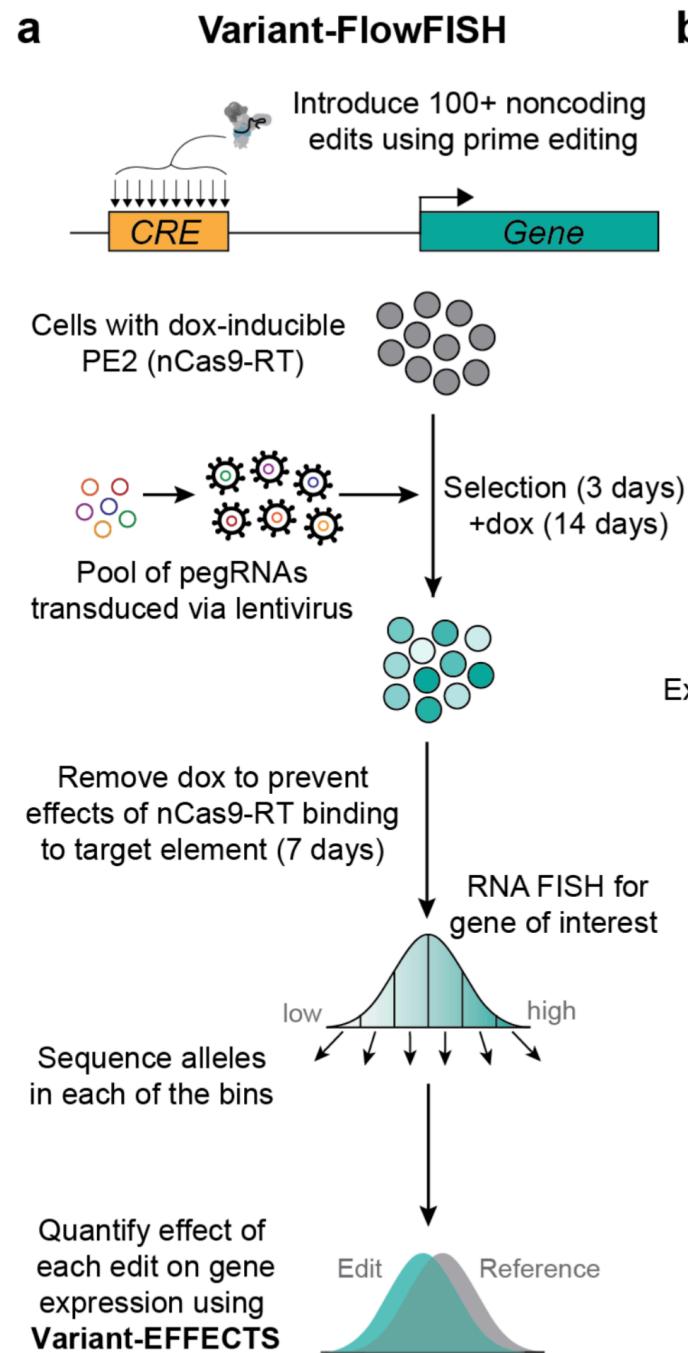
MPRA via saturation genome editing

Findlay et al., 2014, Nature (Shendure Lab)



Variant FlowFISH

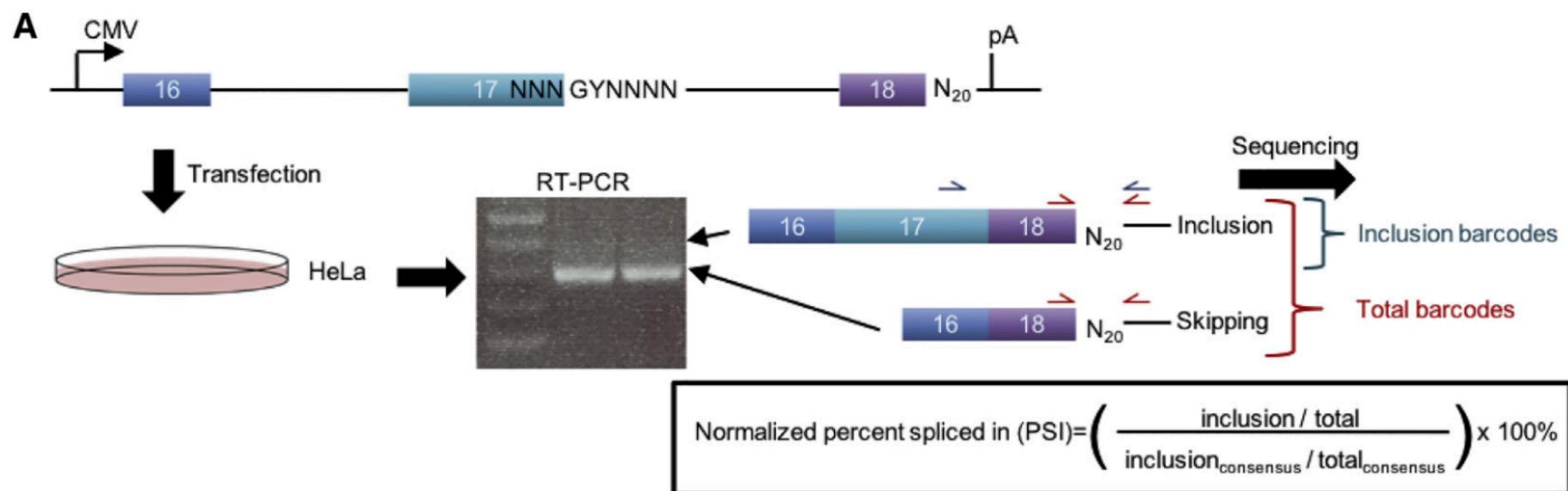
Bartyn et al., 2023, bioRxiv (Engritz Lab)



MPRAs for other aspects of mRNA processing

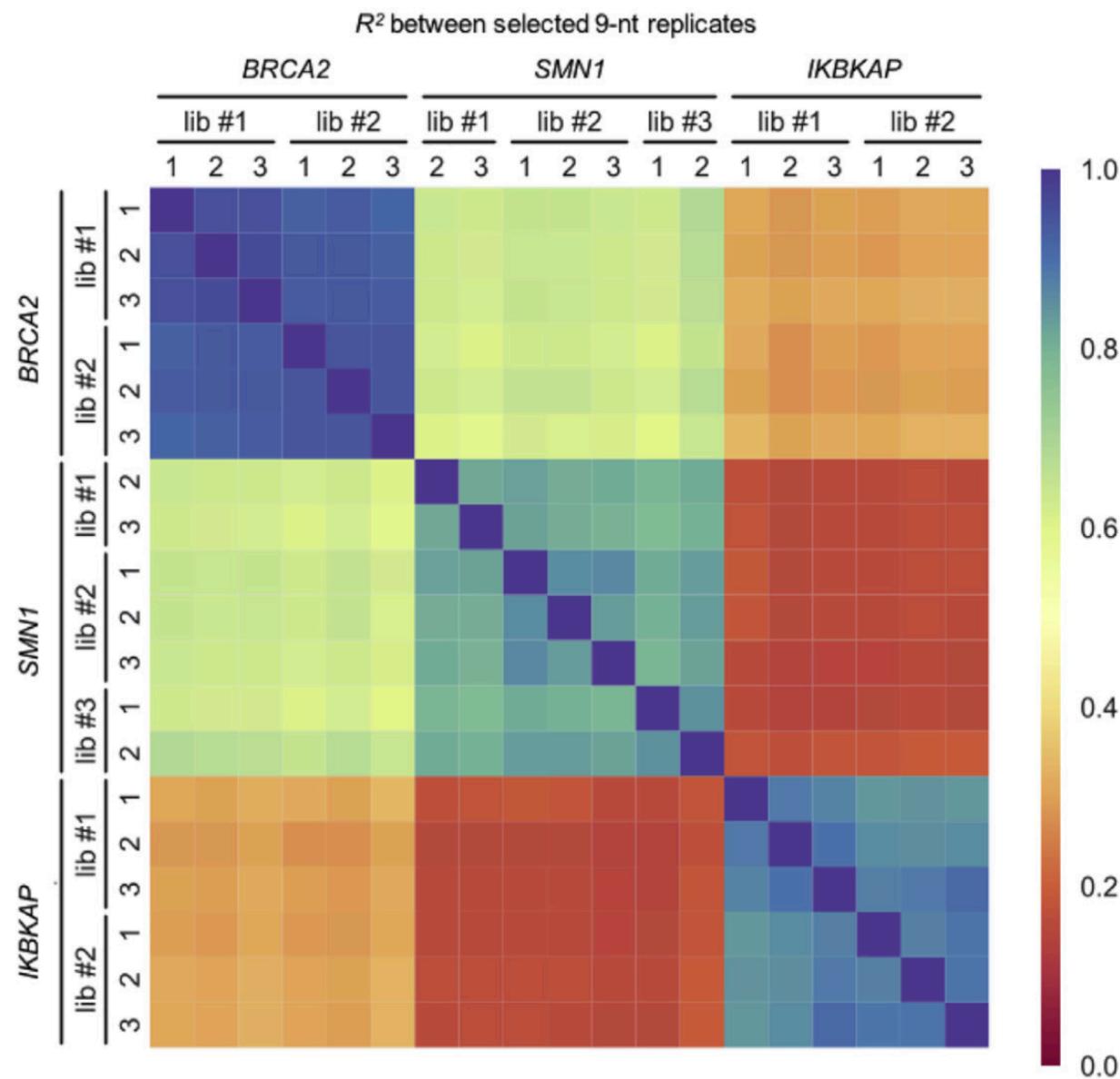
Massively parallel splicing assay

Wong et al., 2018, Mol Cell (Kinney & Krainer Labs)



Massively parallel splicing assay

Wong et al., 2018, Mol Cell (Kinney & Krainer Labs)



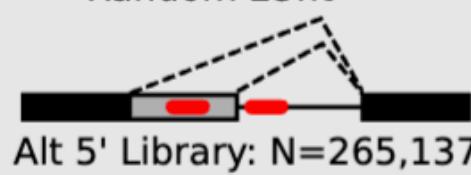
Massively parallel splicing assay

Rosenberg et al., 2015, Cell (Seelig Lab)

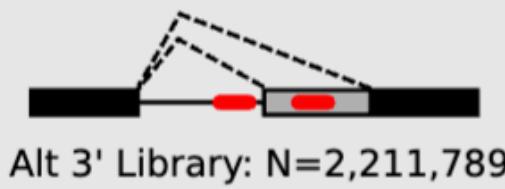
A Learning a Model of Splicing from Millions of Synthetic Sequences

Synthetic DNA Libraries

Random 25nt

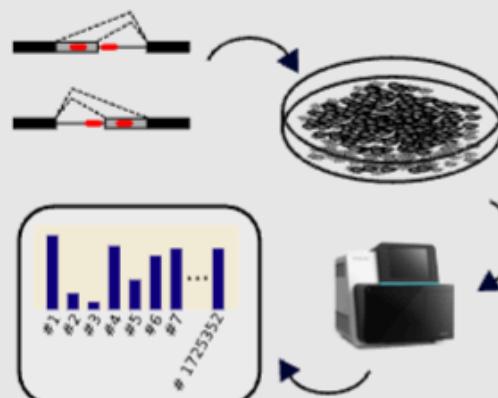


Alt 5' Library: N=265,137

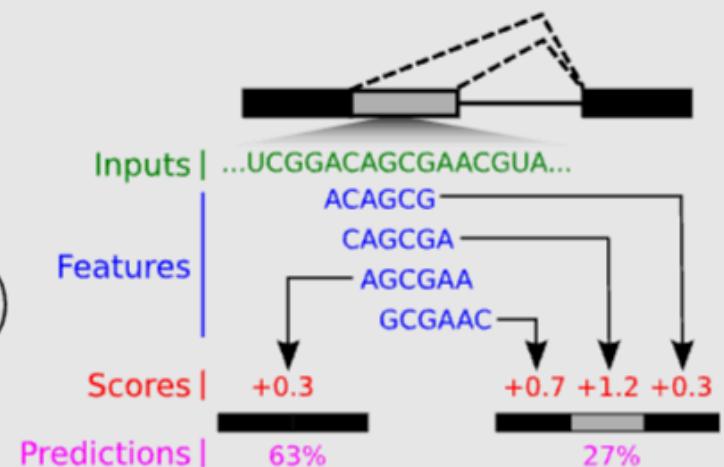


Alt 3' Library: N=2,211,789

Massively Parallel In Vivo Measurement

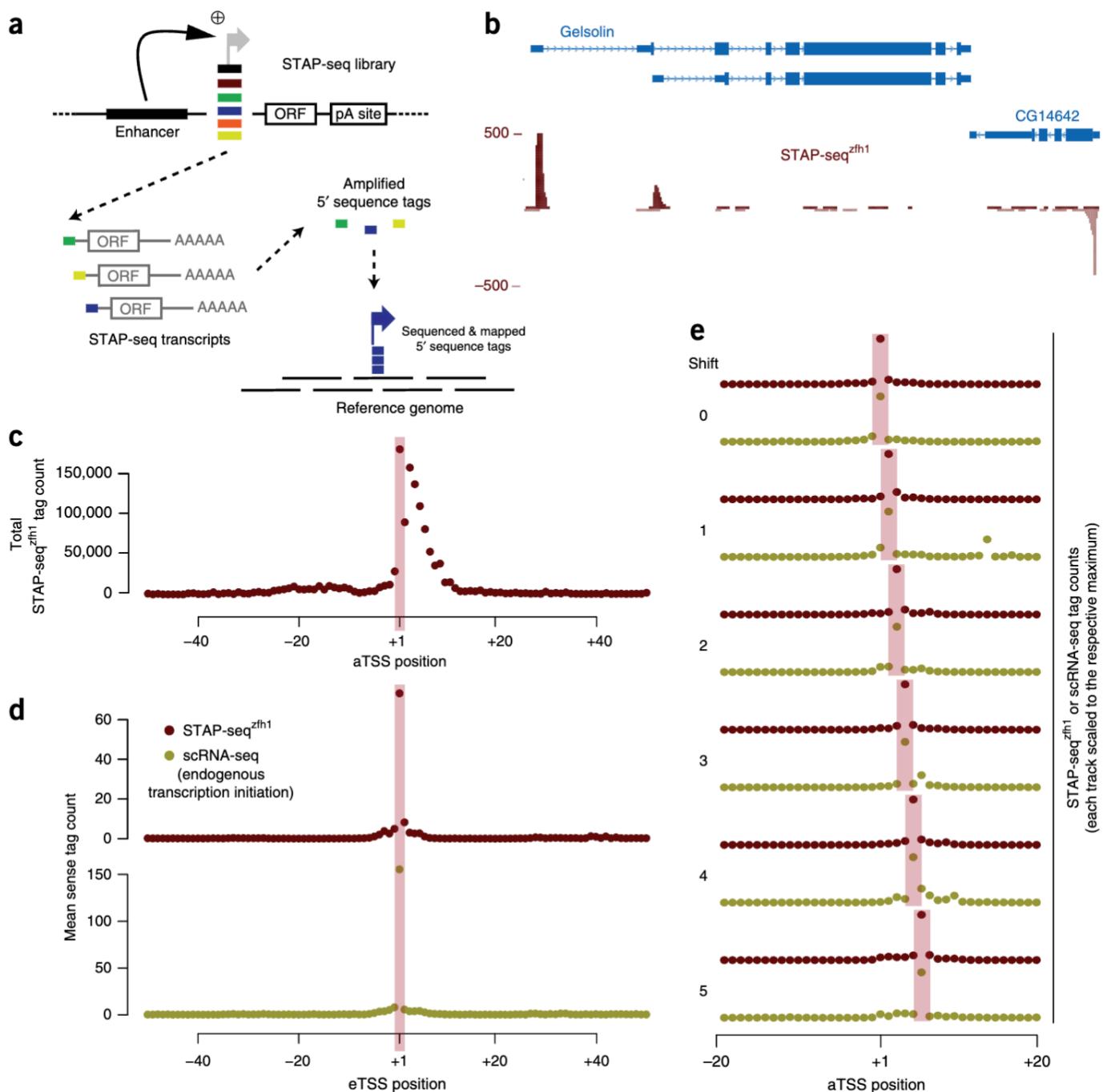


Predictive Model of Splicing



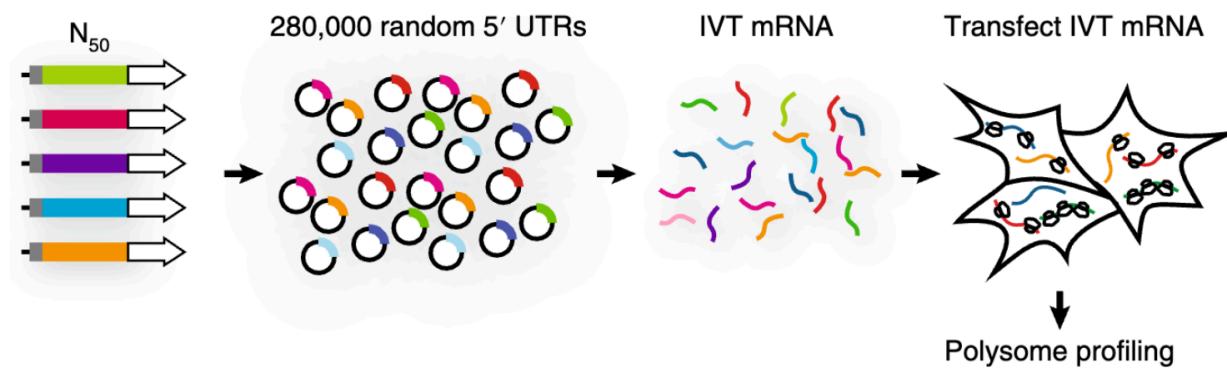
STAP-seq

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

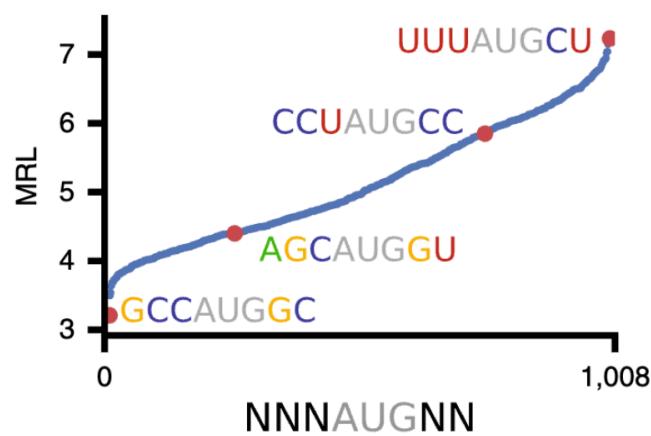


5' UTR MPRA by ribosome profiling

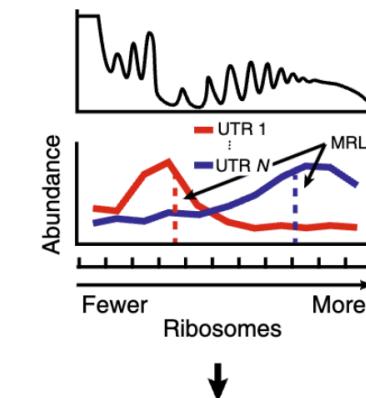
Sample et al., 2019, Nat Biotech (Seelig Lab)



d



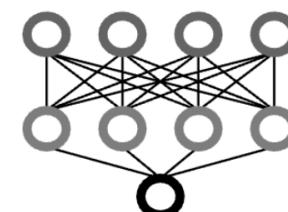
e



Sequence fractions, calculate MRL

- UTR 1: CGGUCACU... MRL: 1.9
- ⋮
- UTR N: AGUUUAGG... MRL: 8

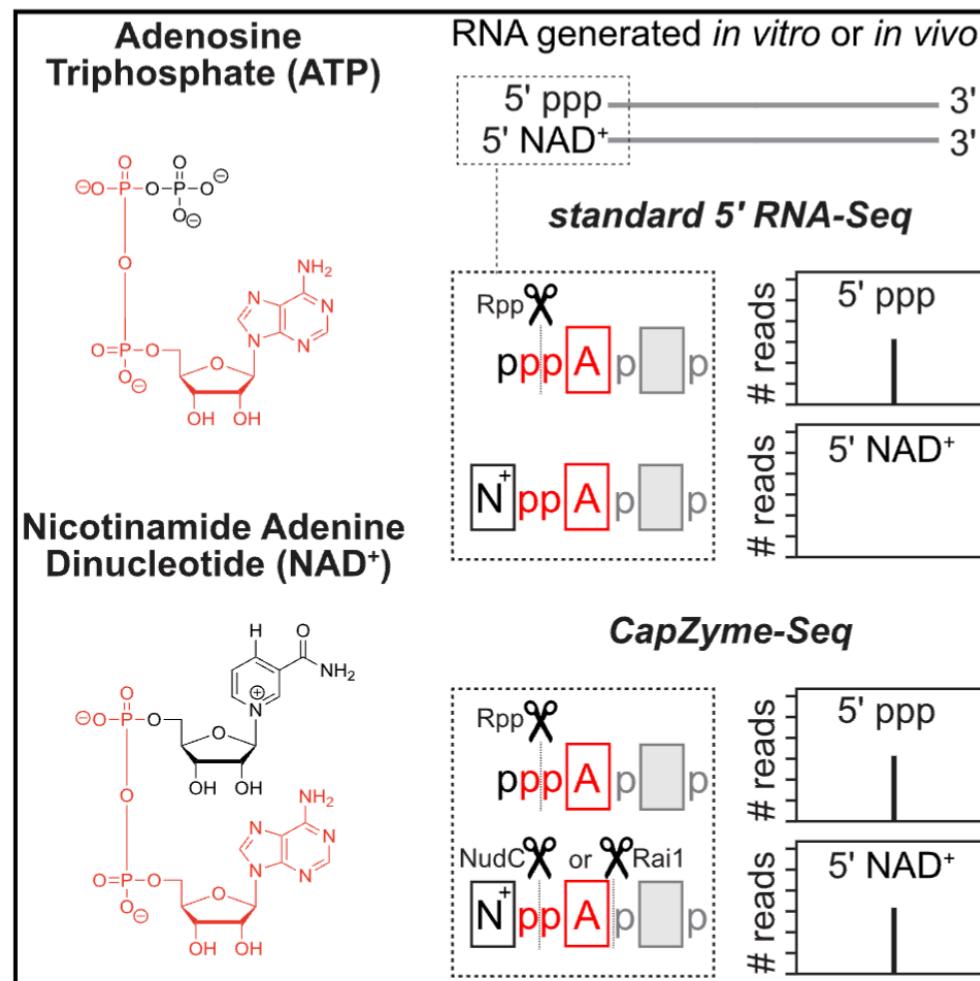
Train CNN



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.)

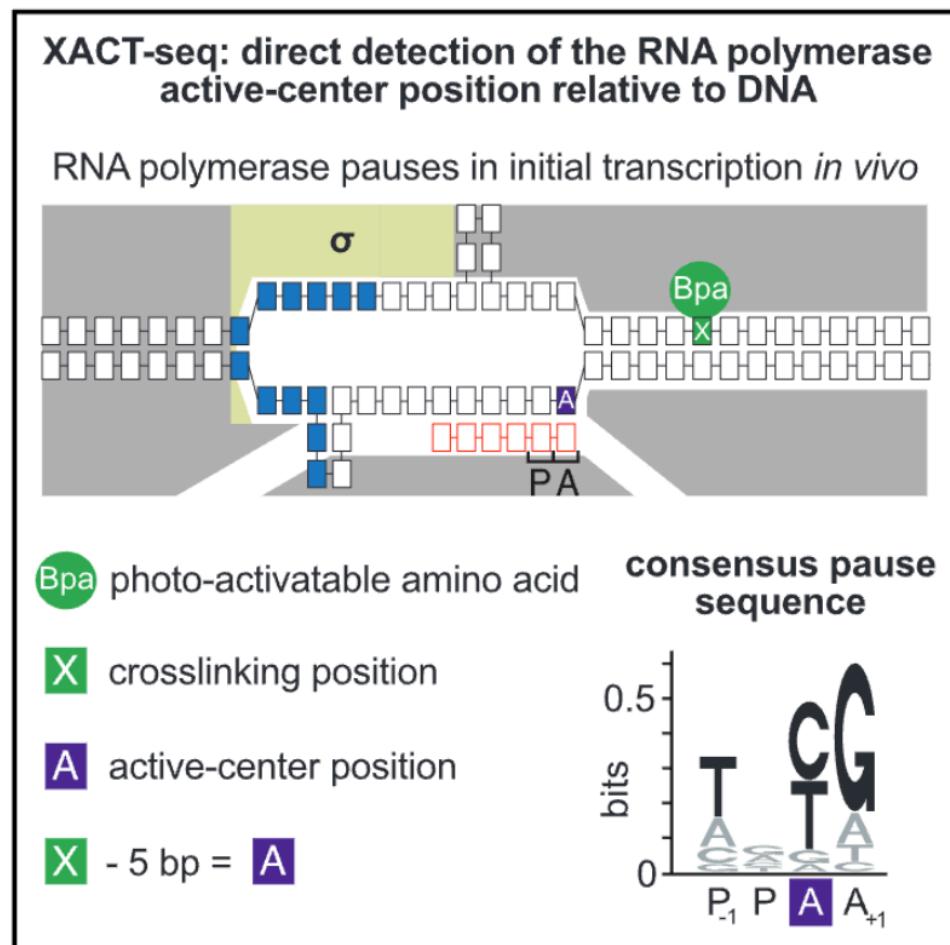
In Brief

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

Winkelman 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

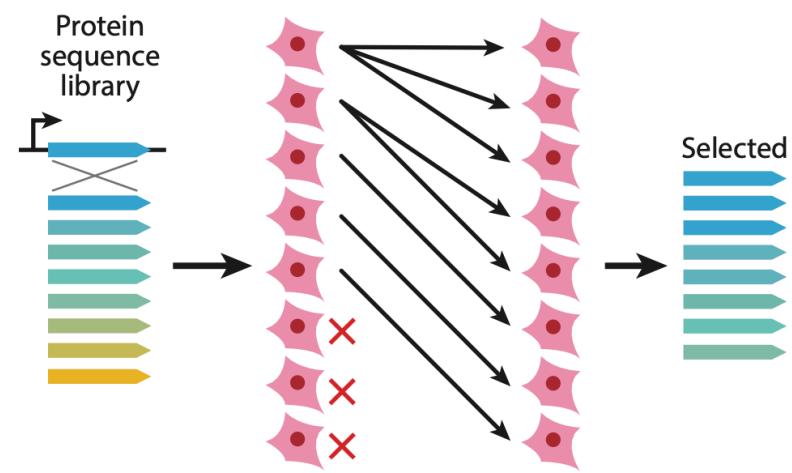
Correspondence

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

In Brief

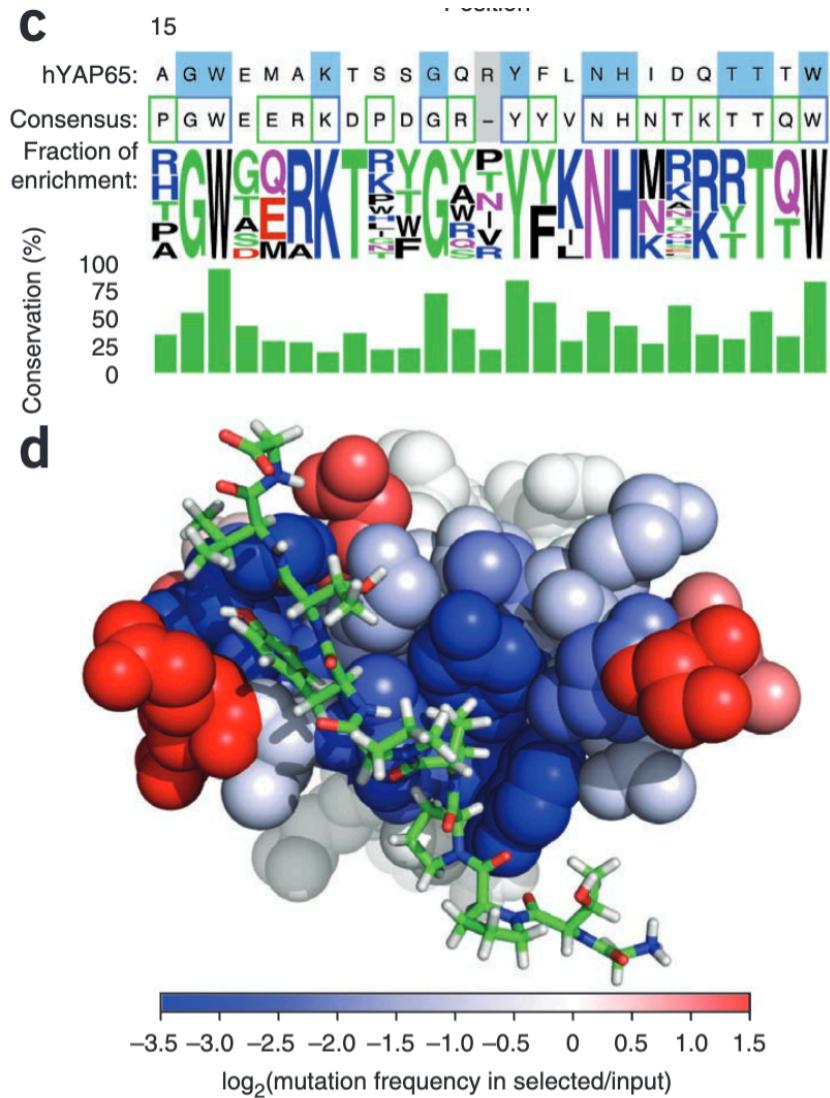
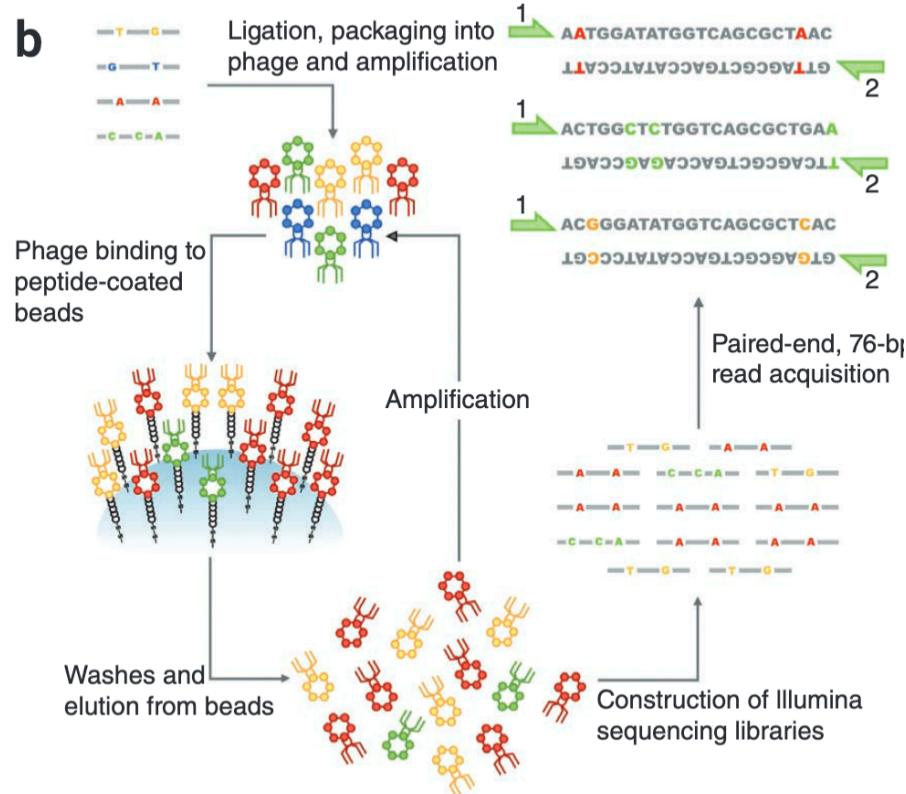
Winkelman et al. report a protein-DNA photocrosslinking method that provides single-nucleotide-resolution readout of RNA polymerase active-center position relative to DNA and enables analysis of initial-transcription pausing. Analysis of 4¹¹ (~4,000,000) promoter sequences defines positional determinants 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



Deep Mutational Scanning (selective growth)

Hietpas et al., 2011, PNAS (Bolon Lab)

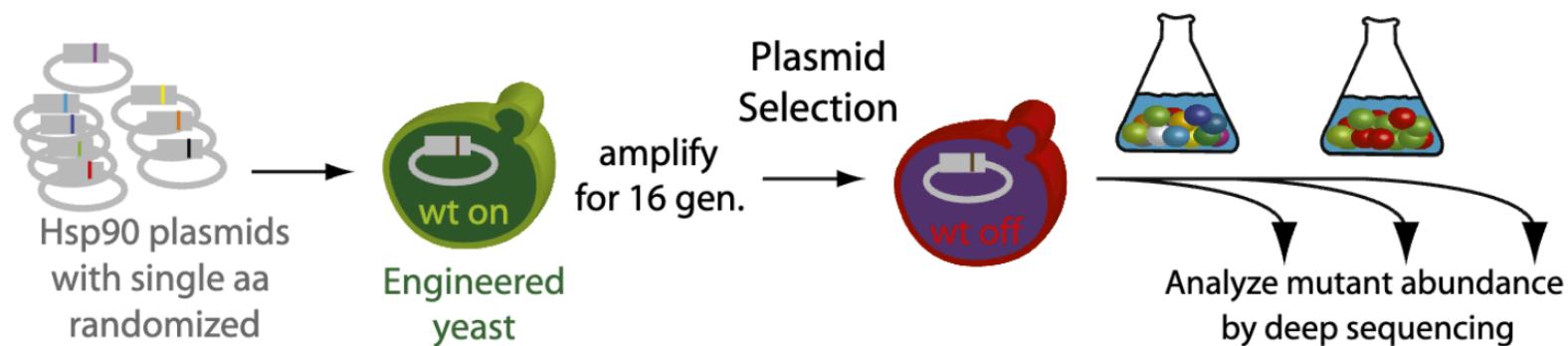
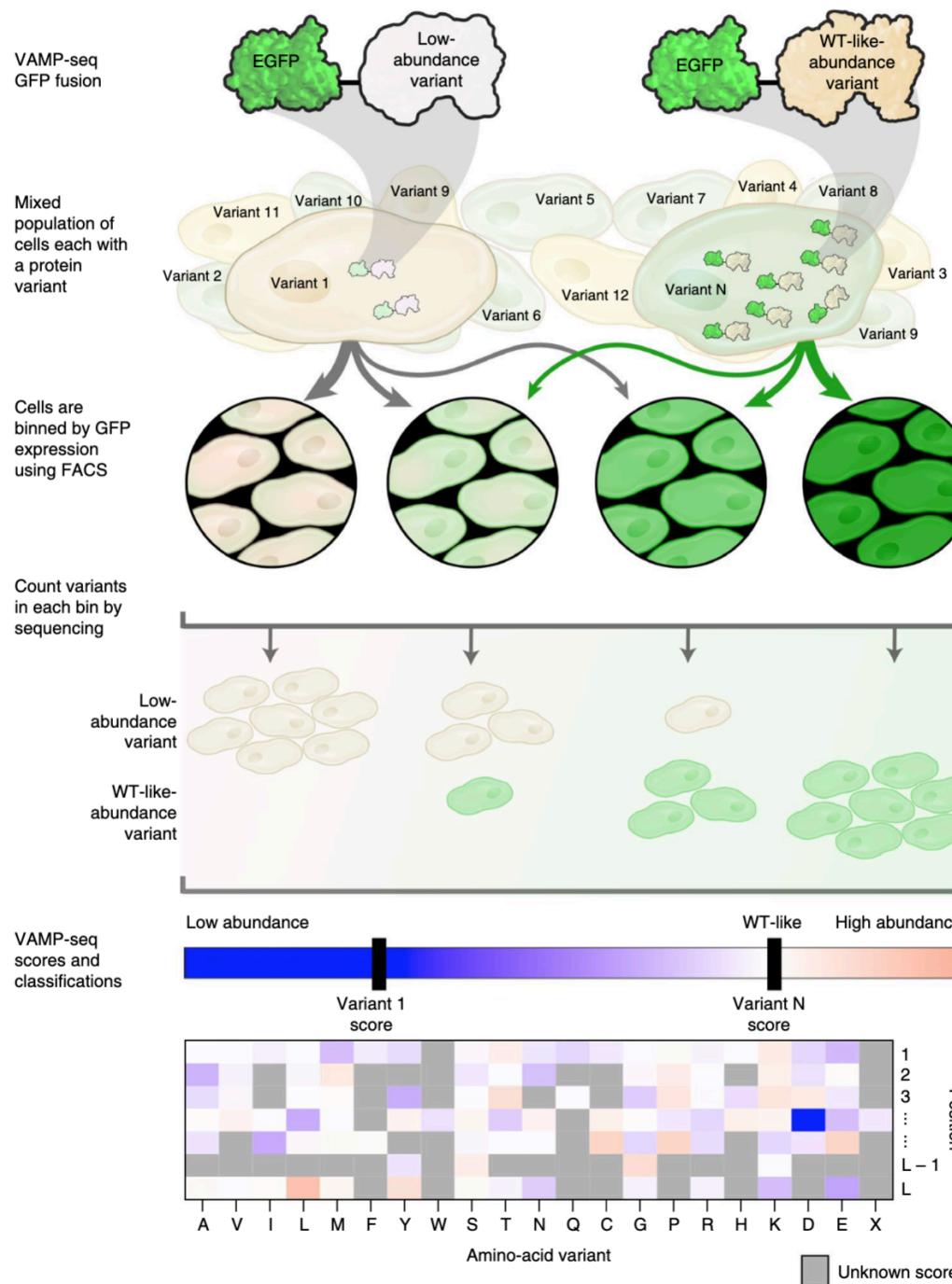
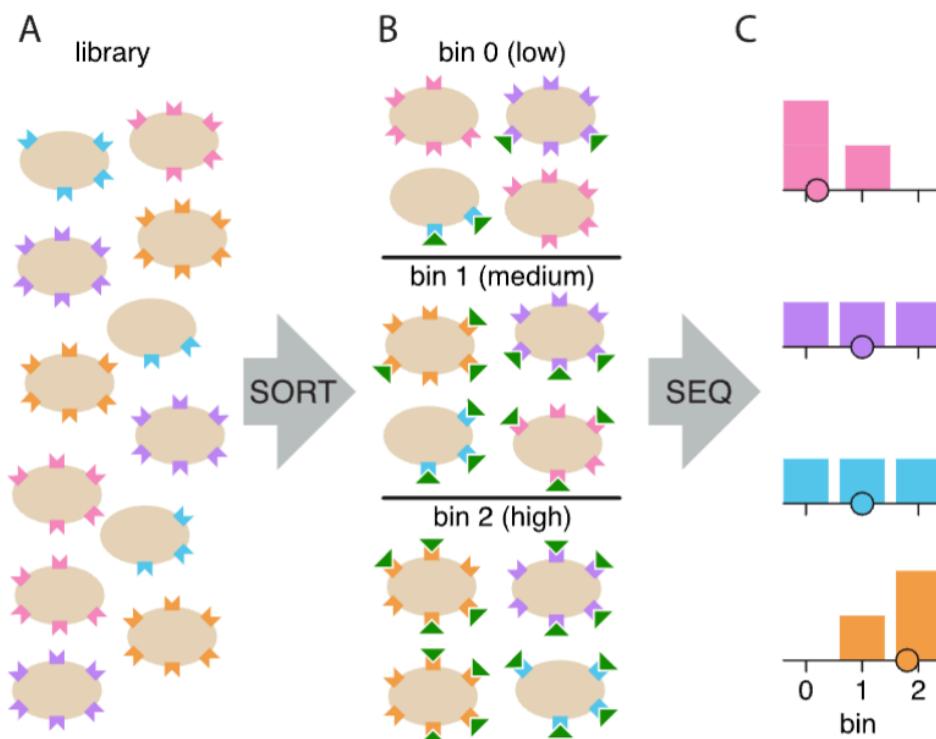


Fig. 1. EMPIRIC approach to experimentally determine fitness landscapes. Randomized individual codon libraries are introduced into a host cell whose only other copy of the gene is regulatable. The fitness of each individual codon mutation is determined by measuring its abundance in the mixed culture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

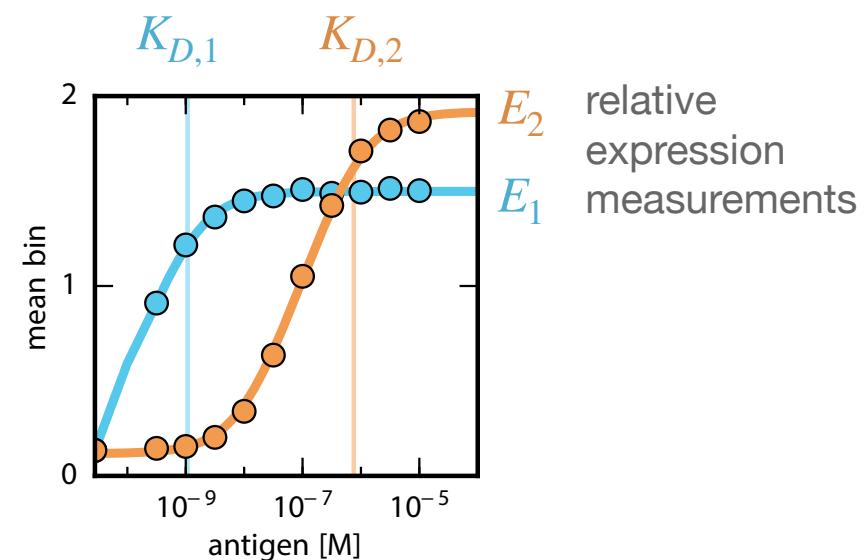
VAMP-seq

Matreyek et al., 2018, Nat Genet



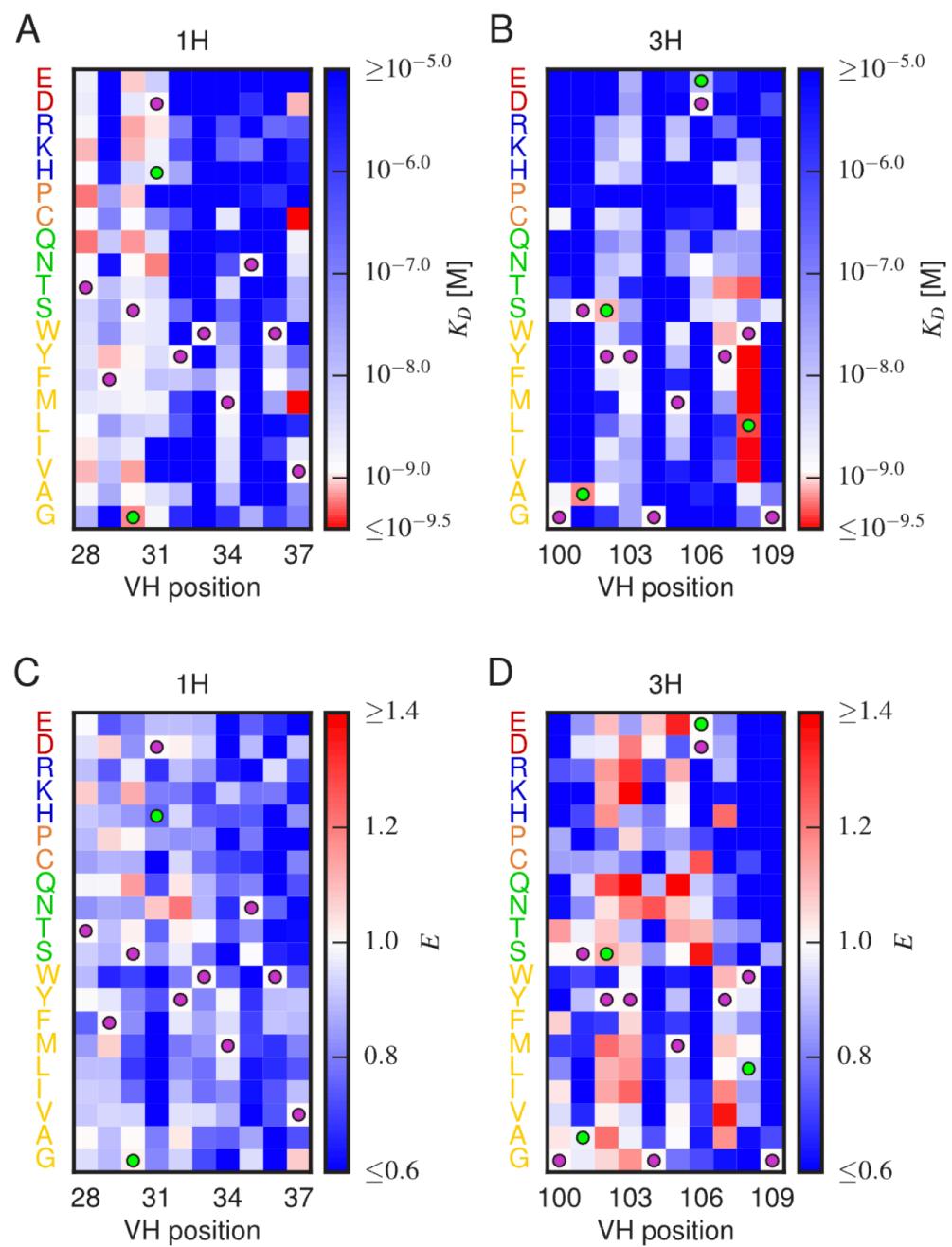
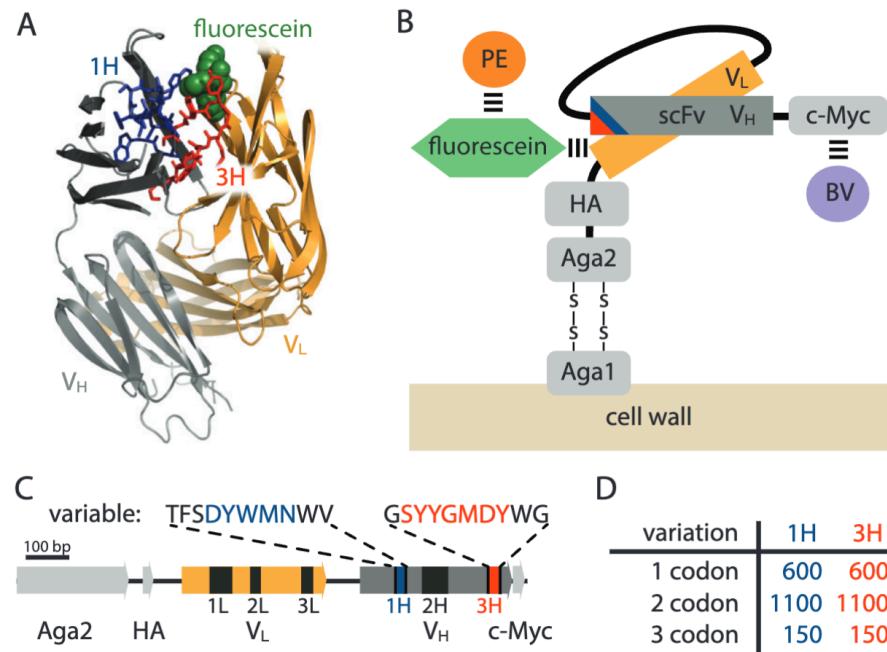


absolute affinity
measurements
(dissociation constants)



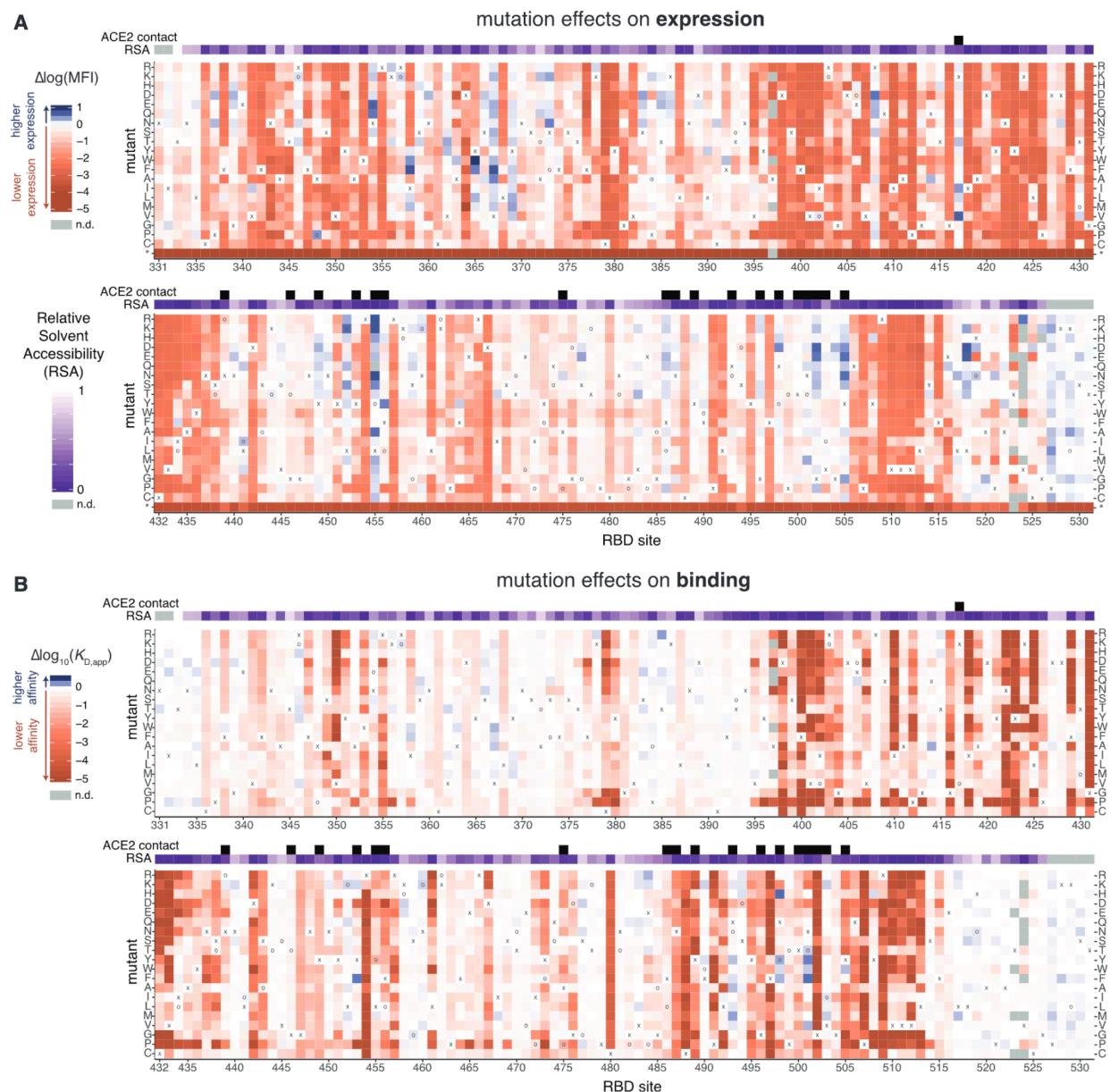
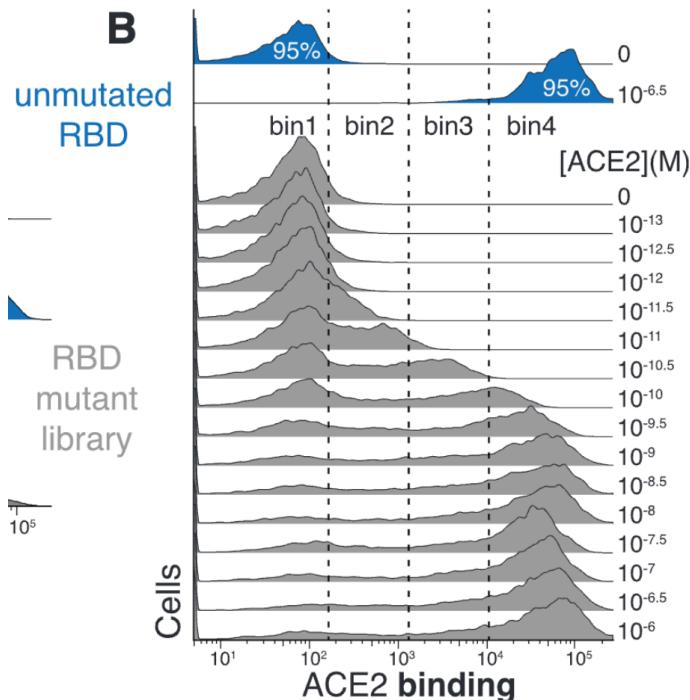
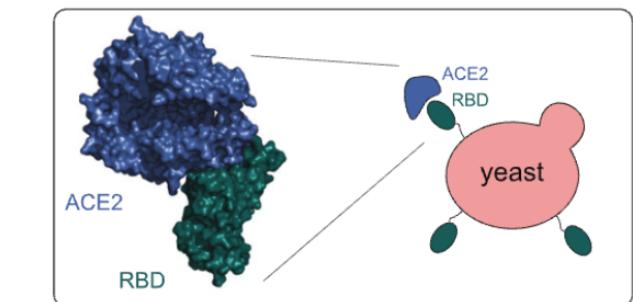
Tite-Seq

Adams et al., 2016, eLife (Kinney Lab)



Tite-Seq applied to SARS-CoV-2

Starr et al., 2020, Cell (Bloom Lab)



ddPCA (doubledeep protein fragment complementation)

Faure et al., 2022, Nature

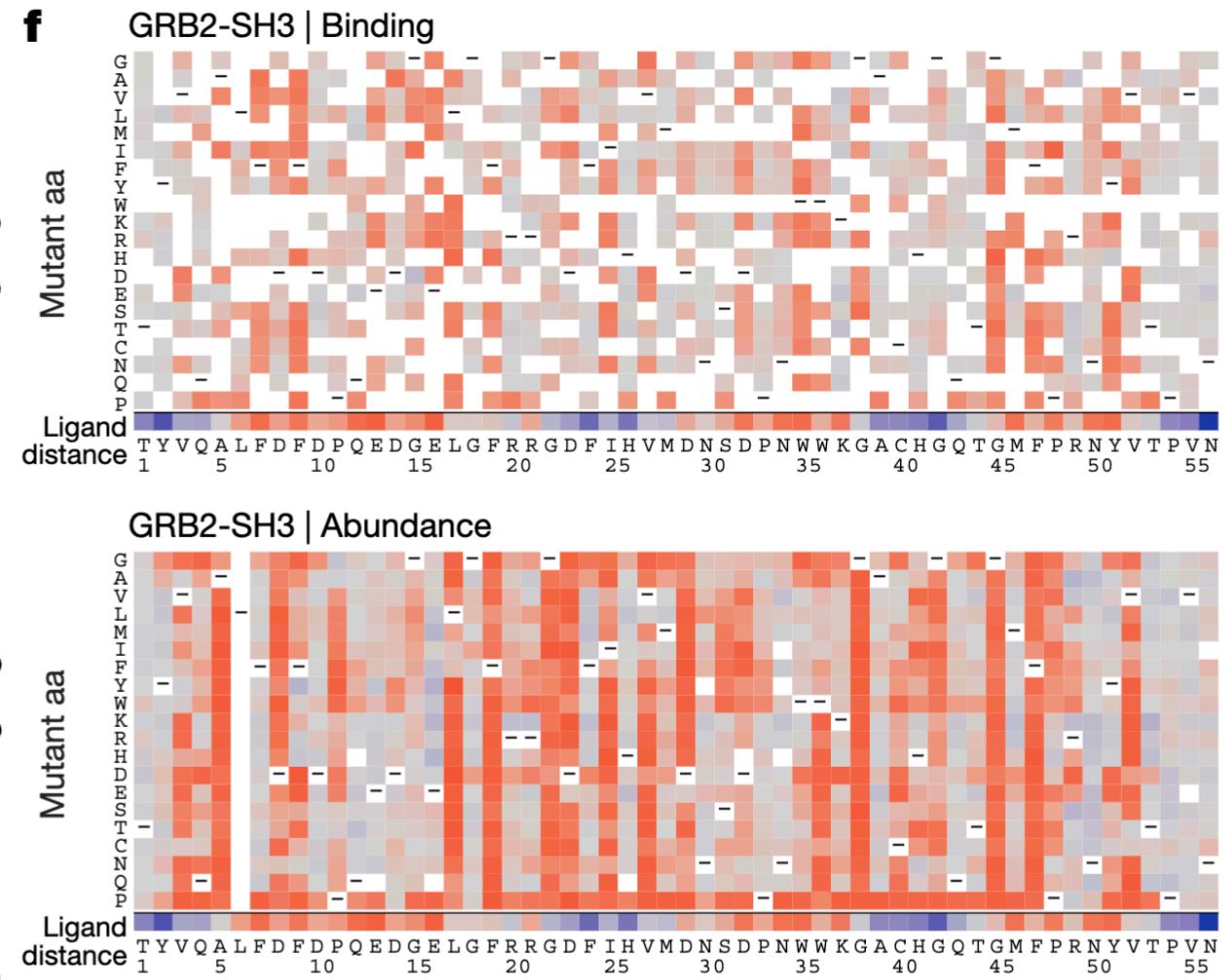
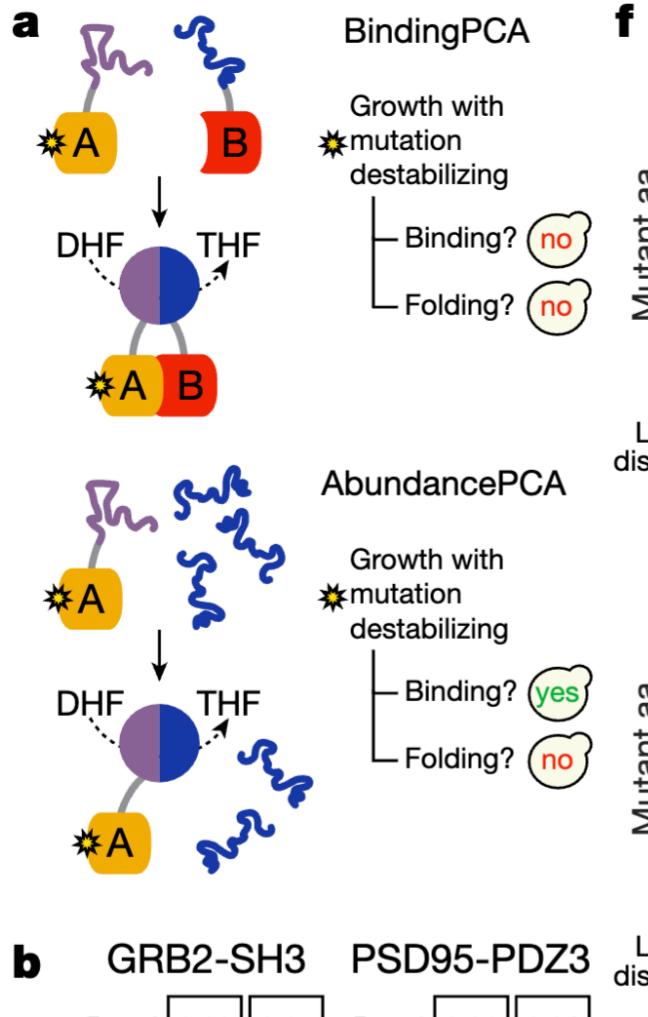
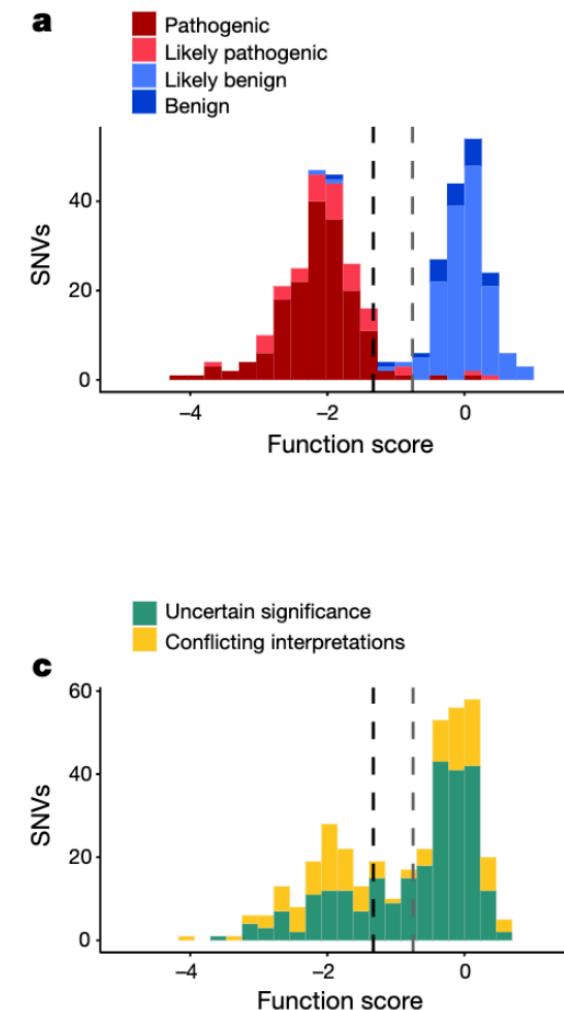
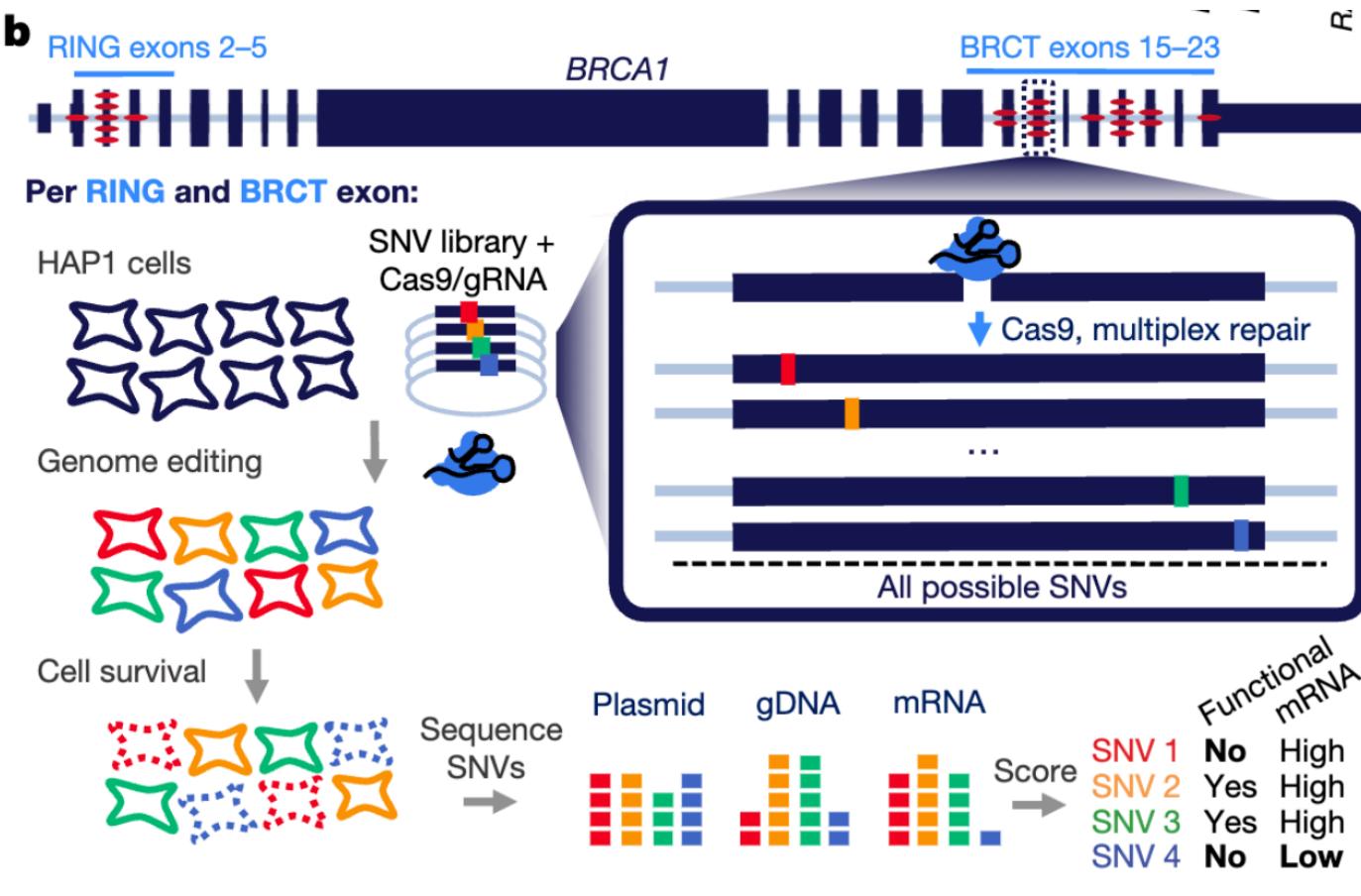


Fig. 1 | ddPCA quantifies the effects of mutations on protein abundance and binding. **a**, Overview of ddPCA on protein of interest A and its interaction partner B. yes, yeast growth; no, yeast growth defect; DHF, dihydrofolate; THF, tetrahydrofolate. **b**, Scatter plots showing the reproducibility of fitness

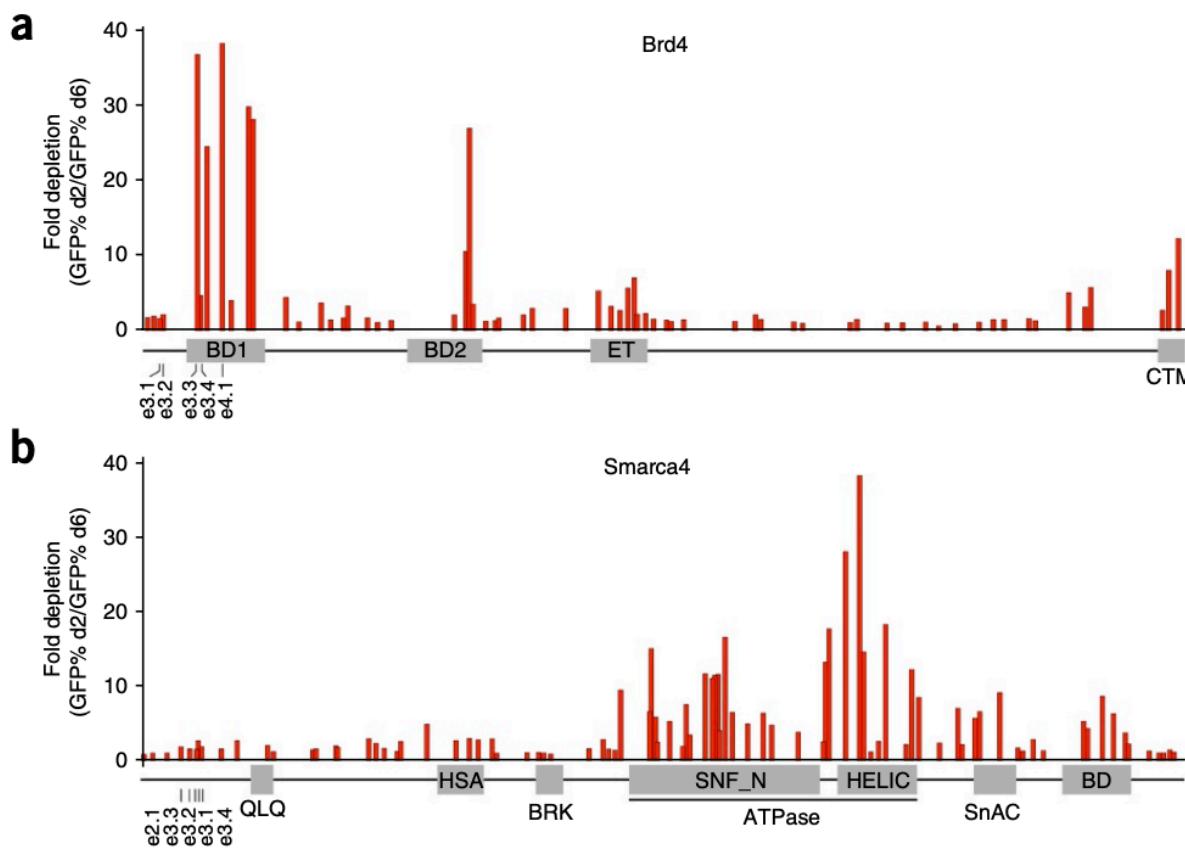
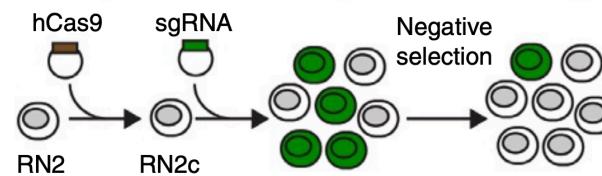
DMS via saturation genome editing

Findlay et al., 2018, Nature (Shendure Lab)



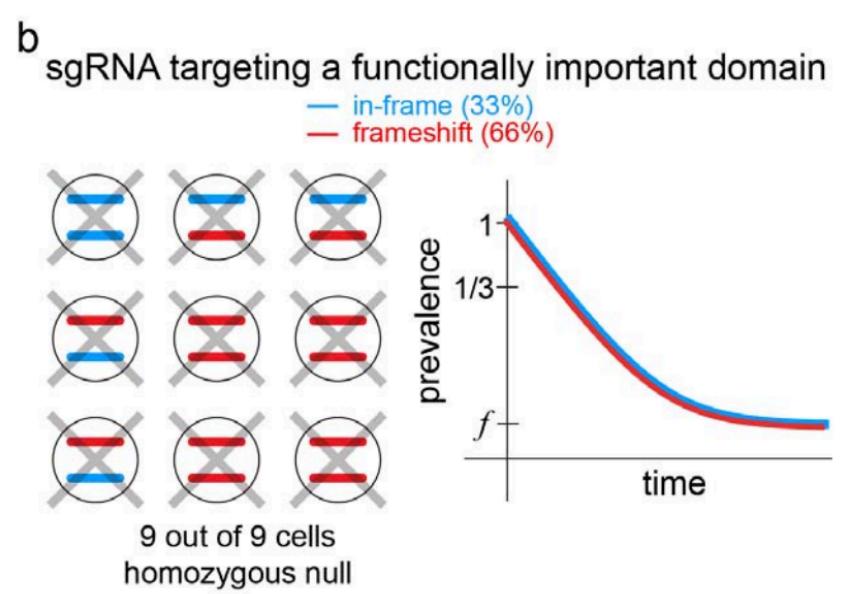
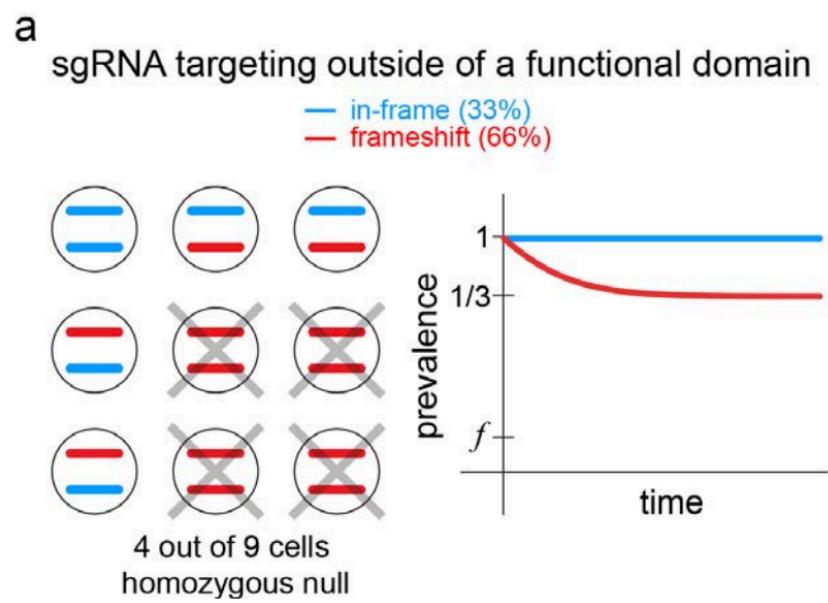
Domain-focused CRISPR screens

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



Search



Home

Search

Documentation

Sign in

About

MaveDB is a public repository for datasets from Multiplexed Assays of Variant Effect (MAVEs), such as those generated by deep mutational scanning (DMS) or massively parallel reporter assay (MPRA) experiments.

MaveDB is open-source, released under the [AGPLv3](#) license.

MaveDB is hosted by the [Fowler Lab](#) in the [Department of Genome Sciences](#) at the University of Washington. It is supported and developed by the [University of Washington](#), the [Walter and Eliza Hall Institute of Medical Research](#), and the [Brotman Baty Institute](#).

Featured Searches

Organisms

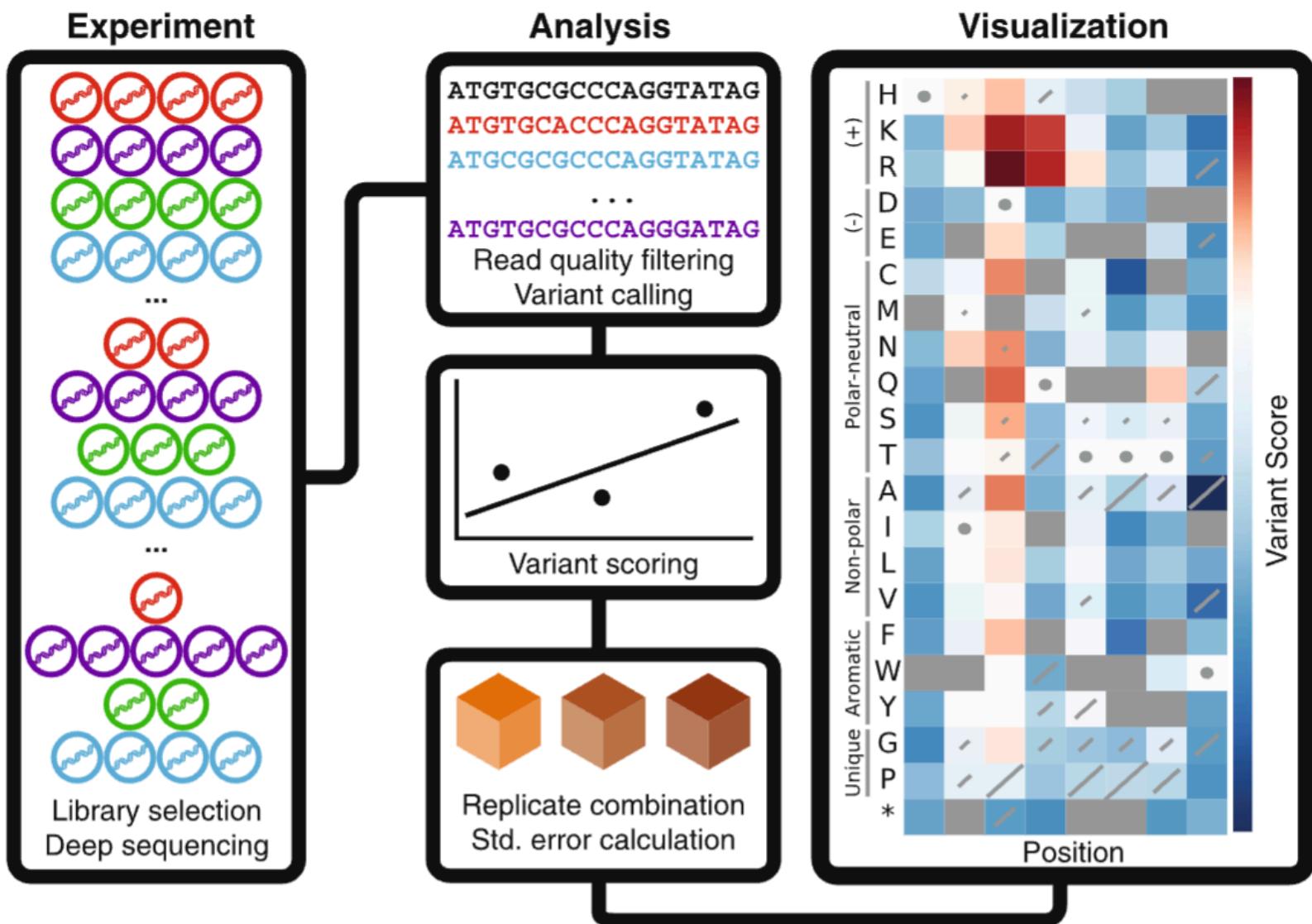
- [Homo sapiens](#)
- [Mus musculus](#)
- [Saccharomyces cerevisiae S288C](#)

Target genes

- [HSP90](#)
- [KCNQ4](#)
- [TEM-1 β-lactamase](#)

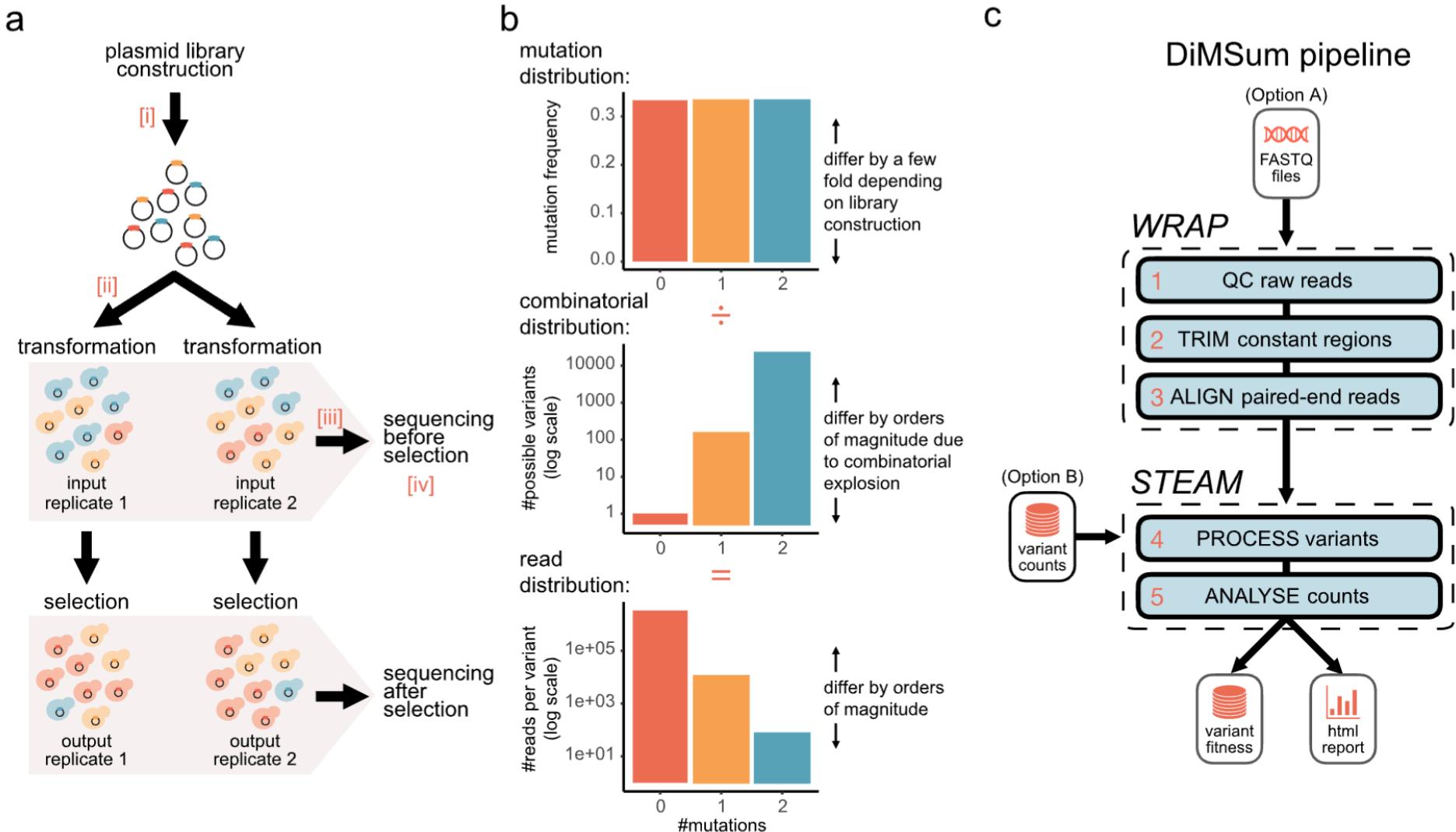
Enrich2: Software for scoring variants in DMS data

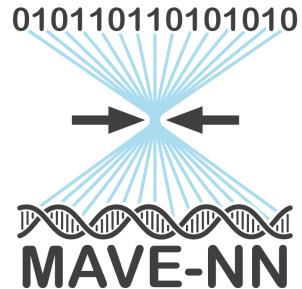
Ruben et al., 2017, Genome Biol (Fowler Lab)



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)

A general computational framework for
quantitatively modeling MAVE data



Ammar Tareen



Anna Posfai

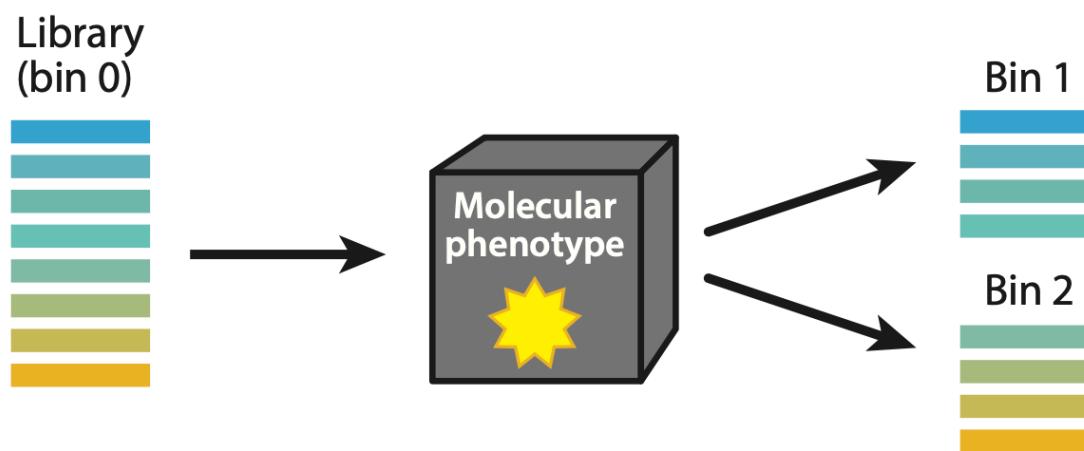


Bill Ireland
(Caltech / Harvard)

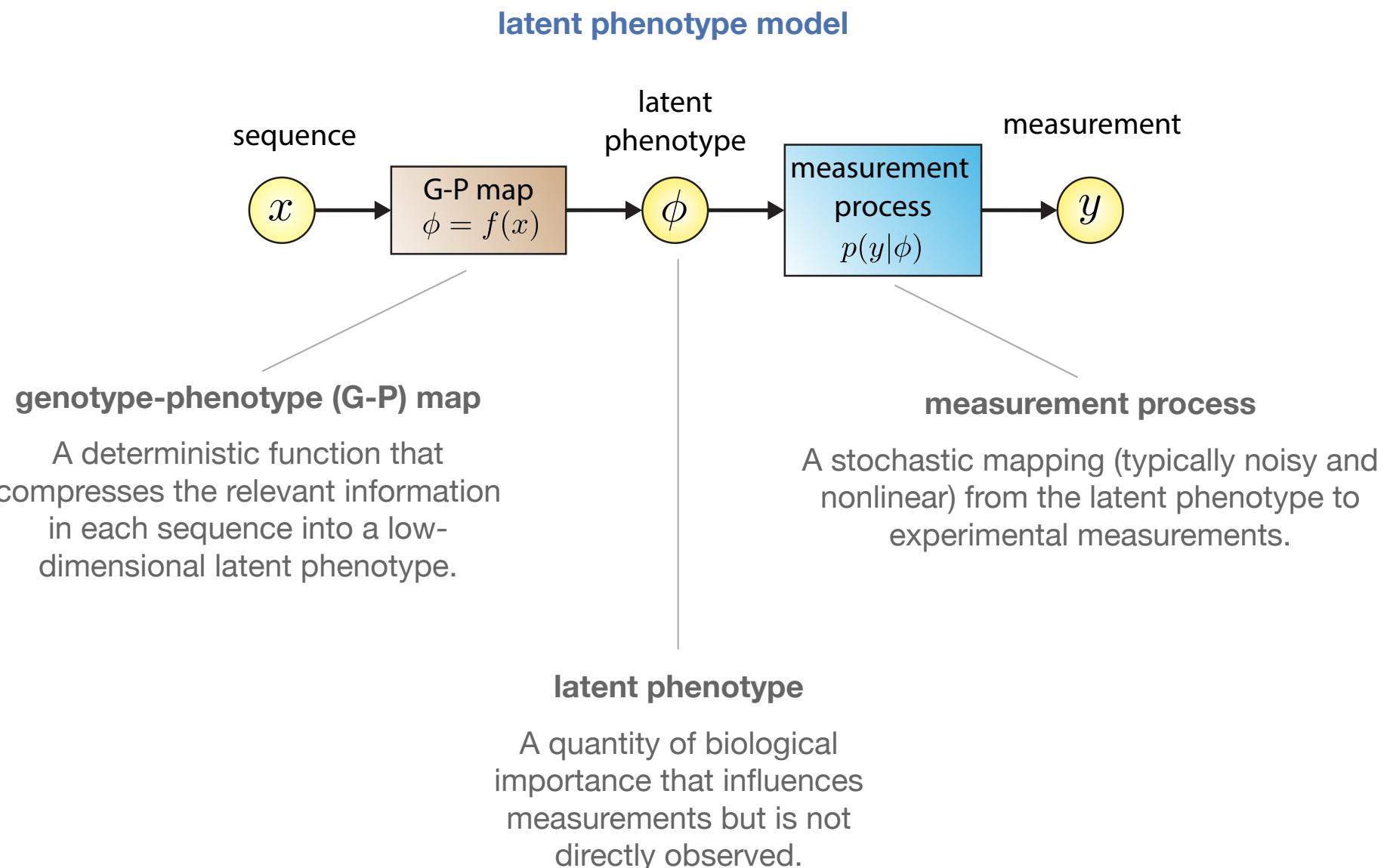


David McCandlish

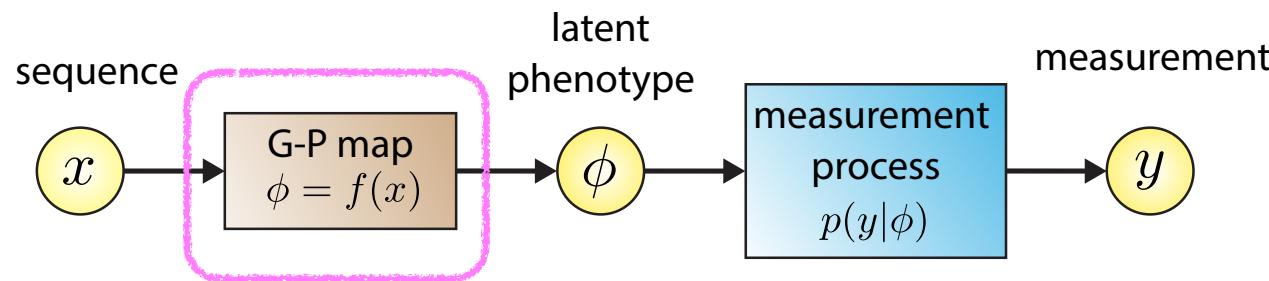
MAVES have a common form that is useful to consider when quantitatively modeling their data.



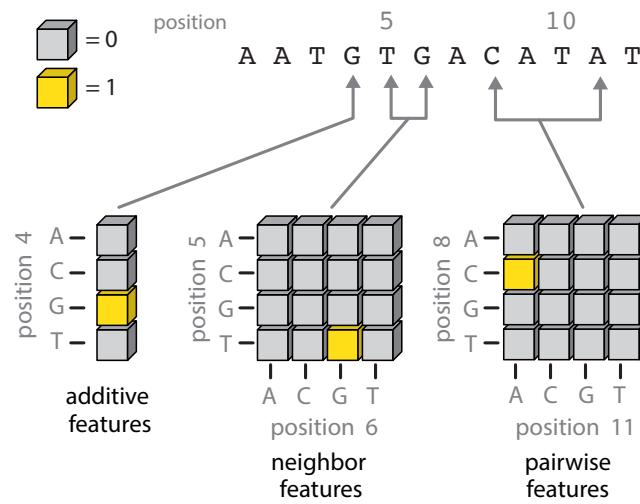
MAVE-NN implements a unified approach for modeling arbitrary genotype-phenotype maps from all types of MAVE experiments.



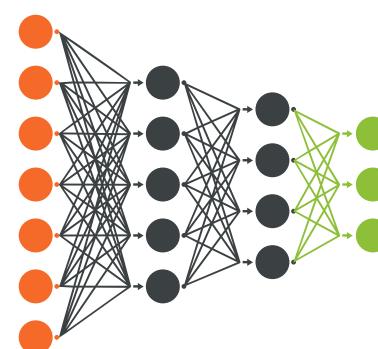
MAVE-NN supports diverse types of models for genotype-phenotype maps.



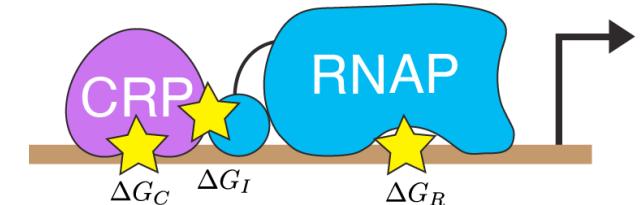
linear feature-based models



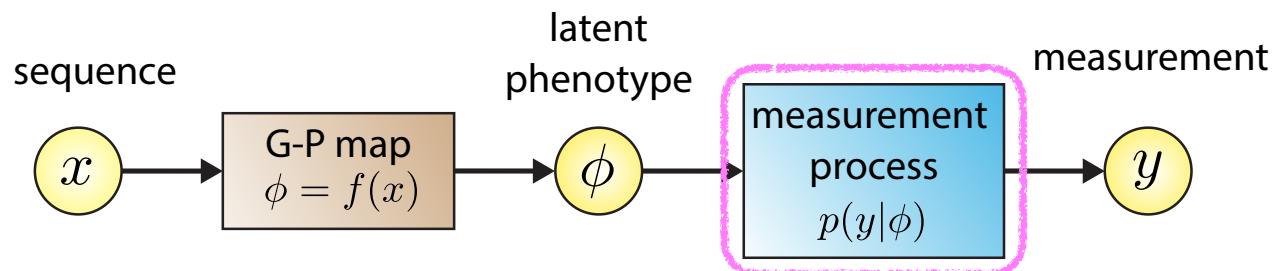
black box neural networks



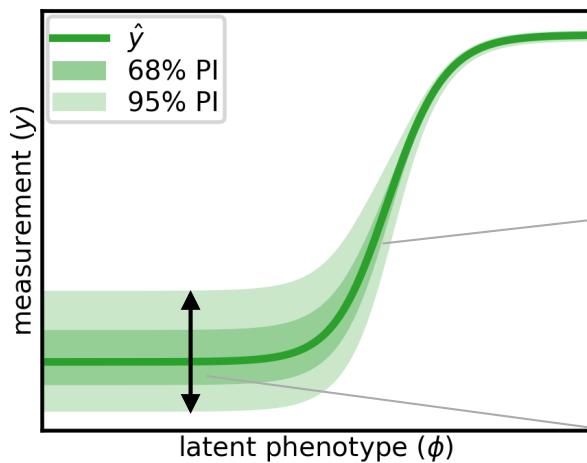
custom genotype-phenotype maps (e.g. biophysical models)



MAVE-NN supports both continuous and discrete experimental measurements.



global epistasis (GE) measurement process

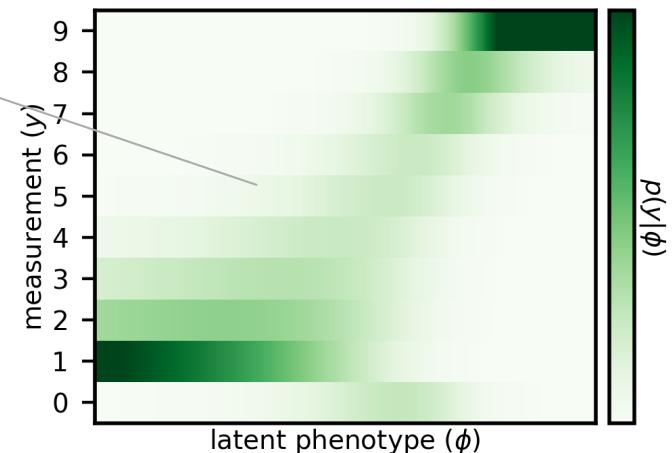


conditional probability
flexible neural network

nonlinearity
sum of sigmoids

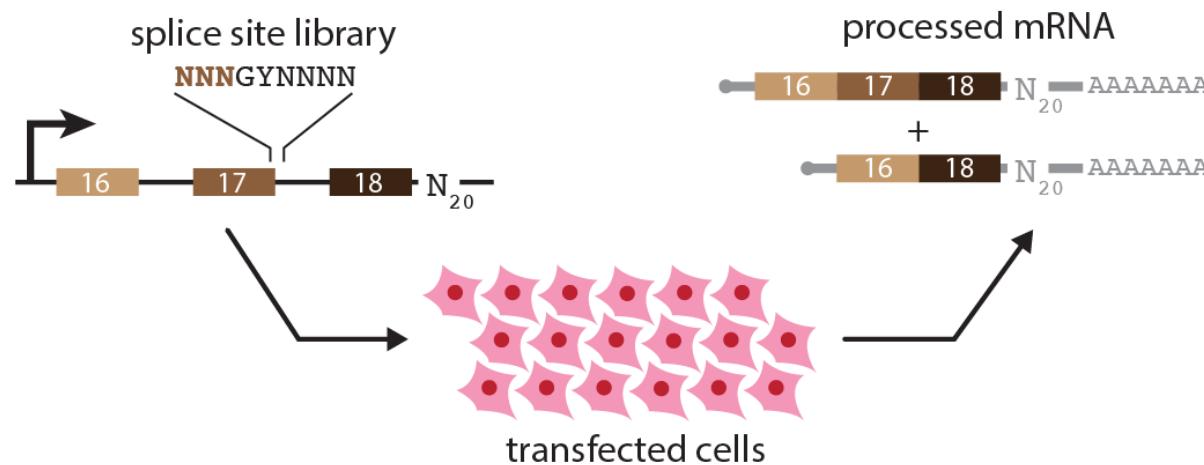
noise model
heteroscedastic
Skewed-t
(Jones & Faddy, 2003)

measurement process agnostic (MPA) measurement process



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%

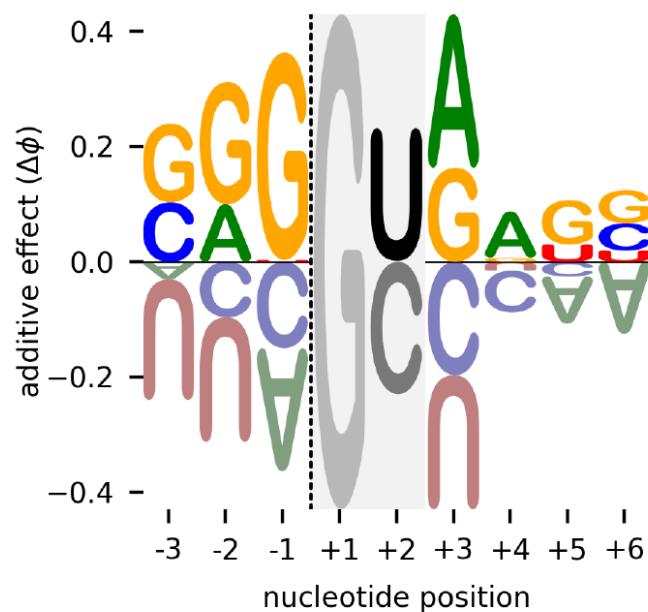
Example: RNA-seq MPRA of 5' splice sites in human cells.

pairwise genotype-phenotype map

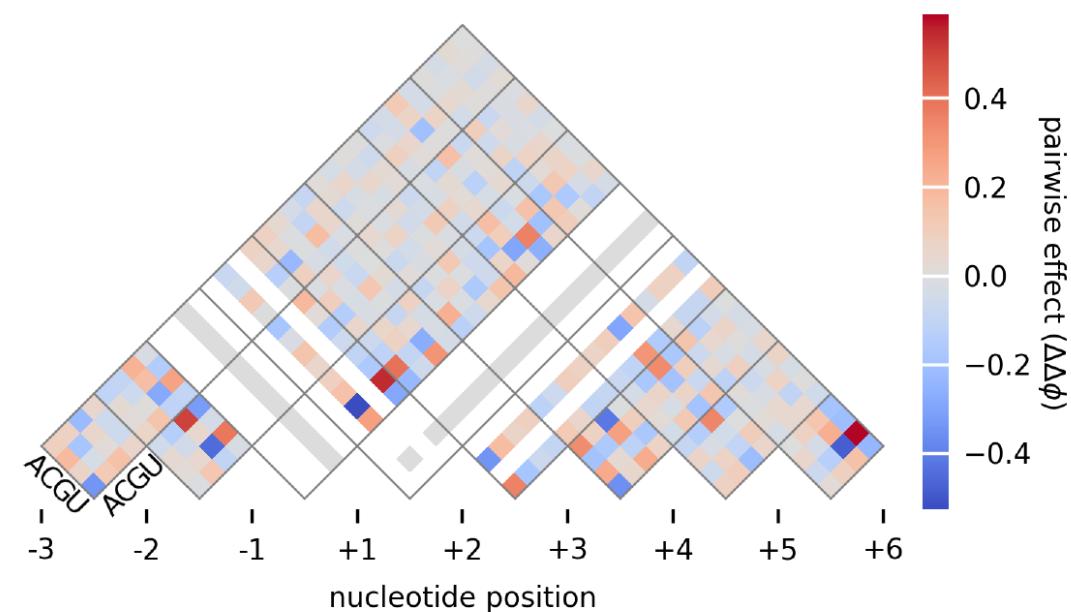
$$\phi(x) = \theta_0 + \sum_{l,c} \theta_{l:c} x_{l:c} + \sum_{l,l',c,c'} \theta_{l:c,l':c'} x_{l:c} x_{l':c'}$$

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

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



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

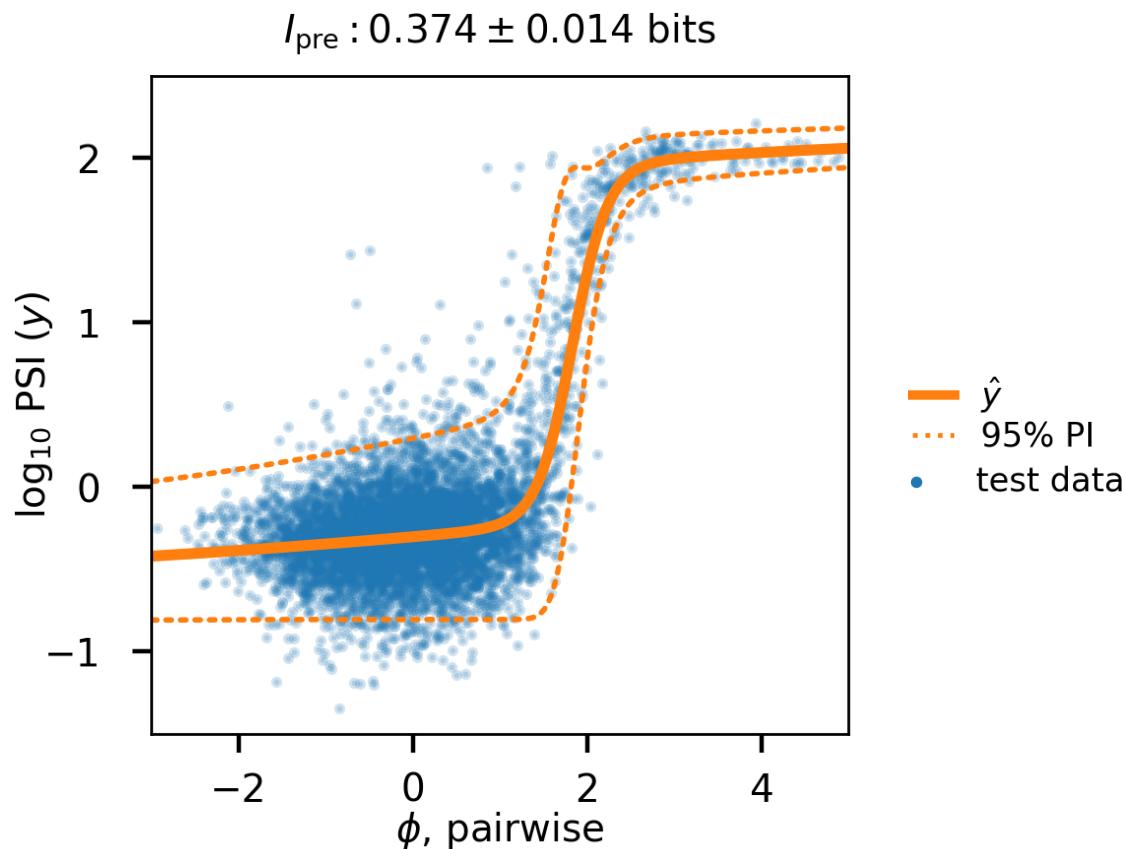


Example: RNA-seq MPRA of 5' splice sites in human cells.

GE measurement process

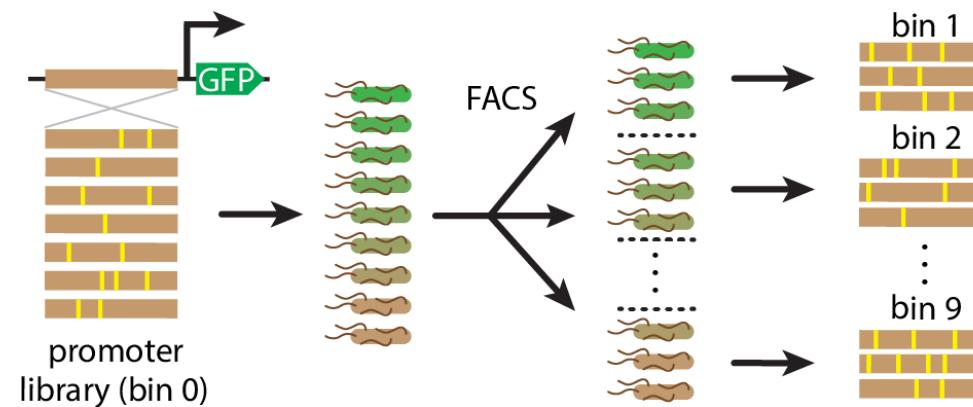
$$p(y | \phi) = \text{SkewedT}(y | \hat{y}(\phi), \eta(\hat{y}))$$

mode shape parameters
/ /



Example: Sort-seq MPRA in bacterial promoters.

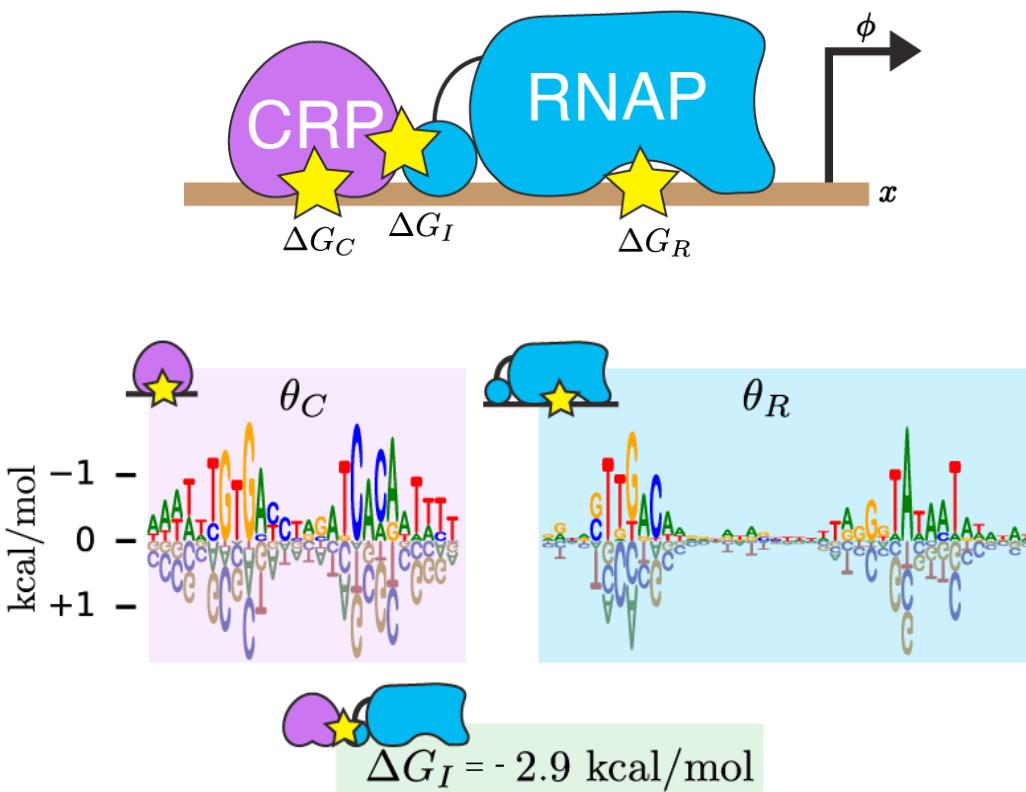
mutagenic study of the *Escherichia coli lac* promoter
(~50K variants sorted into 10 bins)



promoter sequence	read count				
	bin 0	bin 1	...	bin 9	
[yellow mutation]	0	1		0	
[two yellow mutations]	1	0		0	
[one yellow mutation]	0	0		2	
[three yellow mutations]	1	0		1	
[single yellow mutation]	1	0		0	
[four yellow mutations]	0	3		1	
[five yellow mutations]	2	0		0	

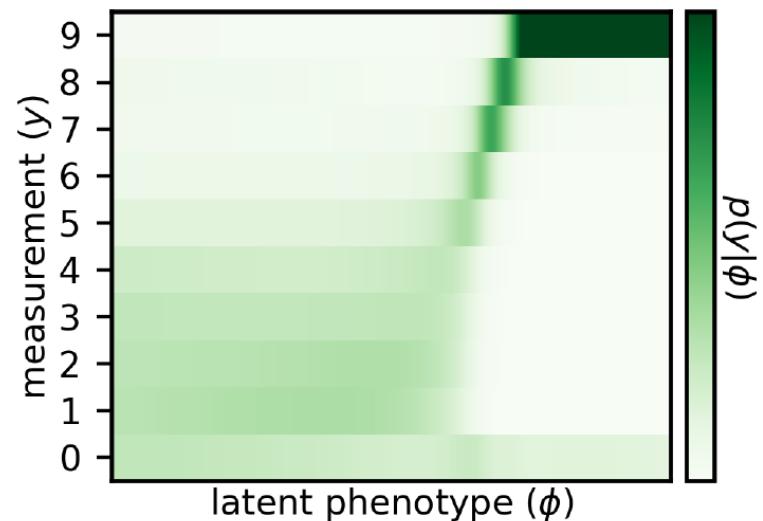
Example: Sort-seq MPRA in bacterial promoters.

thermodynamic genotype-phenotype map



MPA measurement process

inferred measurement process



Kinney 2010: ~1 week on 50 nodes

MAVE-NN: ~15 min on a laptop

Kinney Lab

Lab members

Andalus Ayaz
Debora Tenenbaum
Evan Seitz
Jack Desmarias
Zhihan Liu

Lab alumni

Mahdi Kooshkbaghi
Anna Posfai
Ammar Tareen
Wei-Chia Chen
Mandy Wong
Rhys Adams
Talitha Forcier
Manraj Gill
Kush Koshic
Daniel Jones

Honorary lab members

Yuma Ishigami
(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 (PI)
NIH R01HG011787 (PI)
NIH R01HG012131
NIH P30CA045508
Northwell Health
Moore Foundation

