Discovery with deep learning, human genetics, and perturb-seq

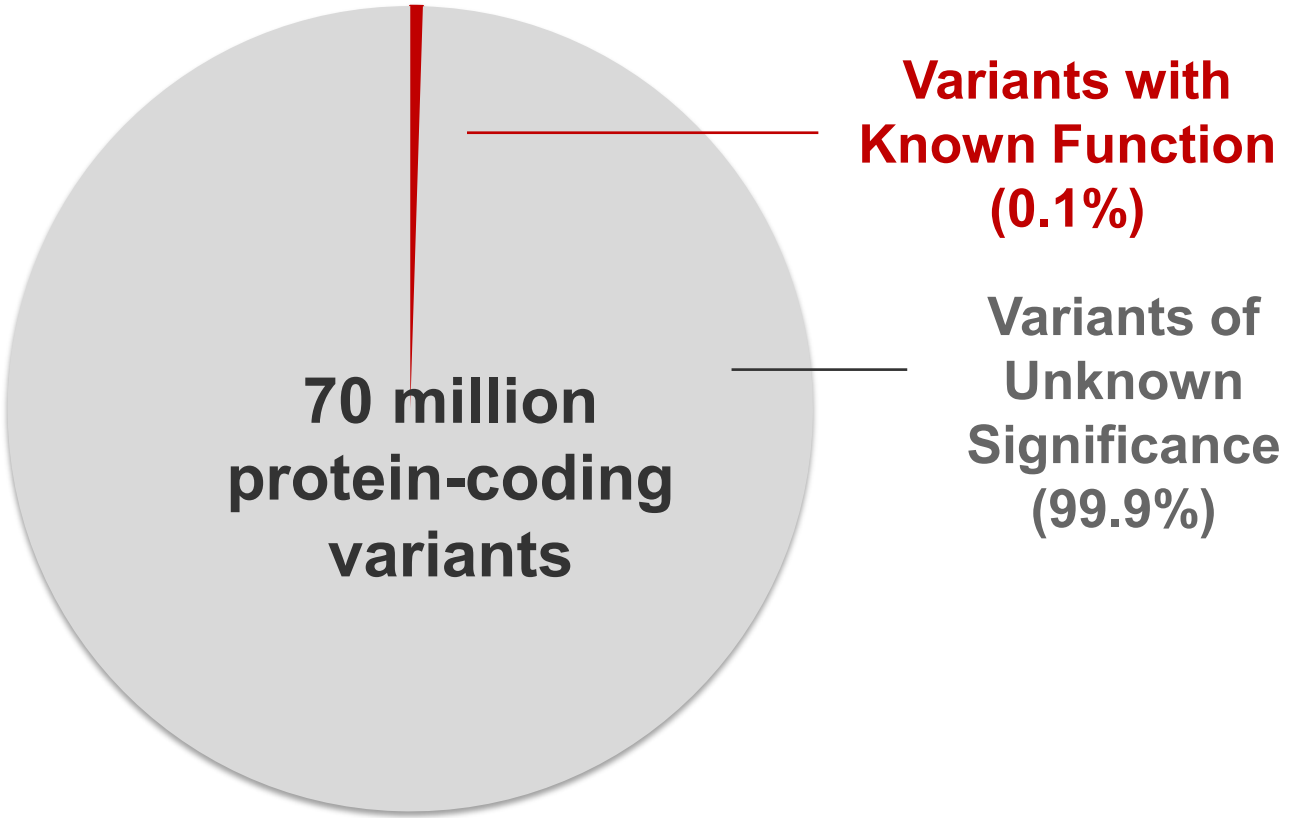
Kyle Kai-How Farh, MD, PhD

VP & Distinguished Scientist, Artificial Intelligence

illumina®

Our 5-year plan: to decipher the effect of all variants in the human genome

Our current knowledge of the clinical effects of genetic variants is nascent

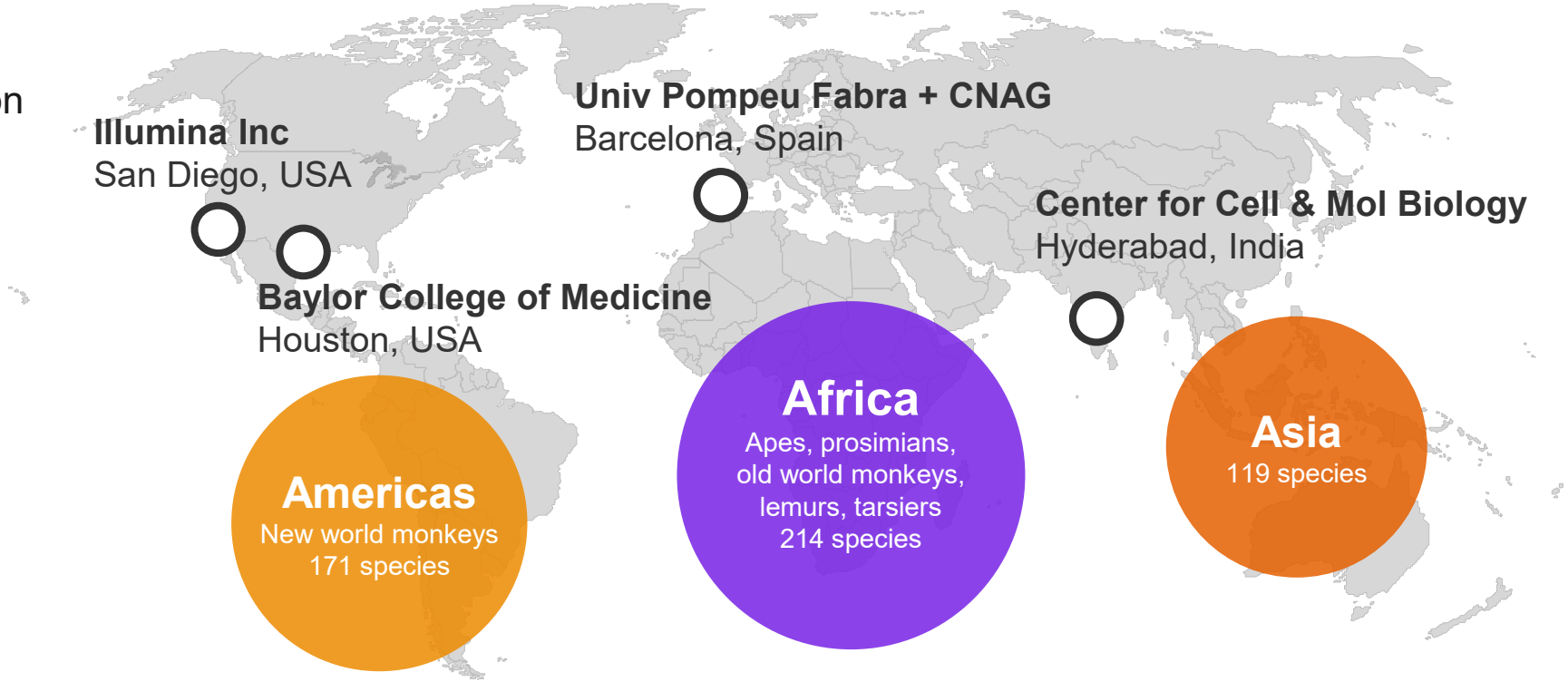


illumina®

PrimateAI-3D resolves variants of unknown significance (VUS) by leveraging large-scale evolutionary data

The Primate Conservation Sequencing Initiative

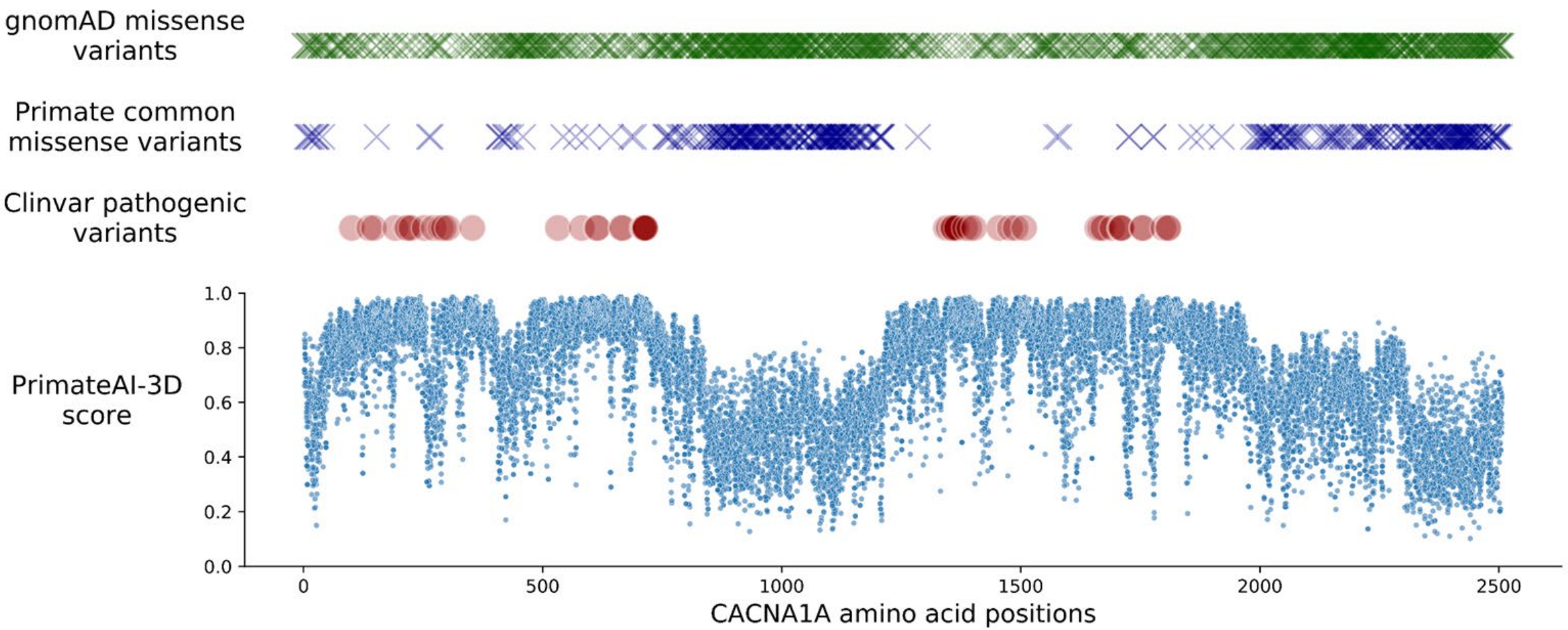
- Deep neural network trained on sequencing of 233 primate species to predict pathogenic mutations in humans
- ~4.4 million human VUS re-classified as likely benign to improve variant interpretation
- 70X larger training data than existing ClinVar database



illumina®

Fiziev et al, Science 2023; Gao et al, Science 2023

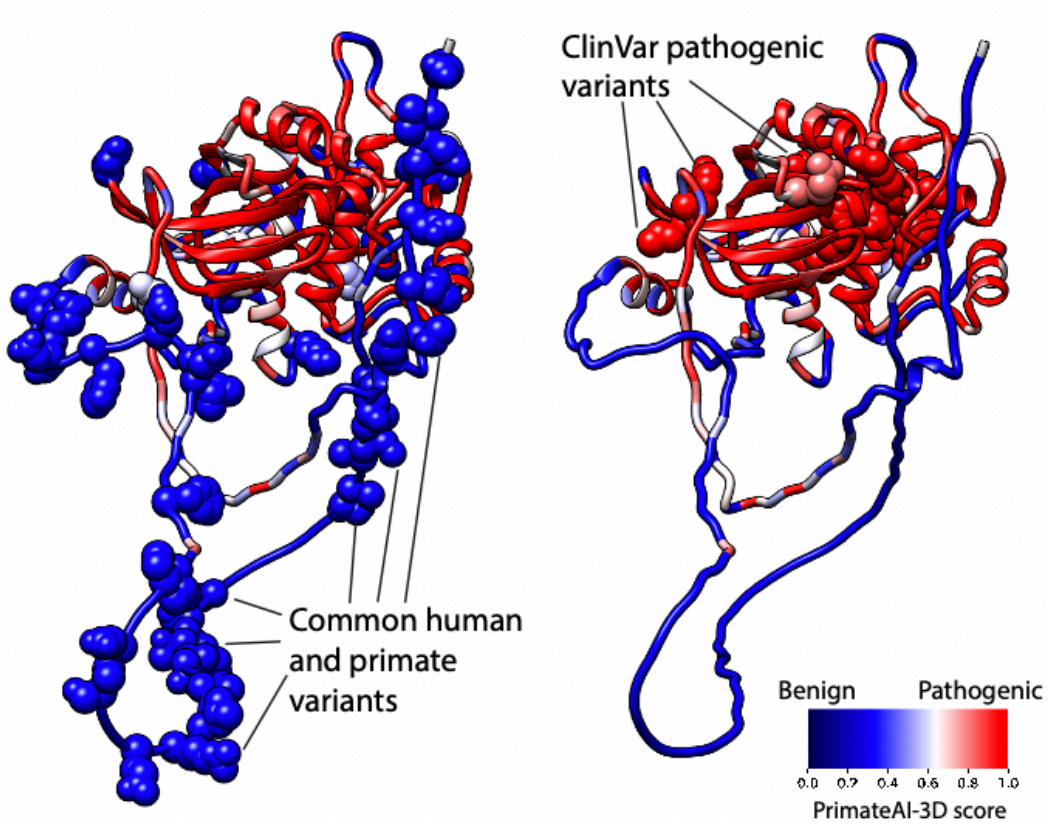
CACNA1A – AA position of primate and ClinVar variants



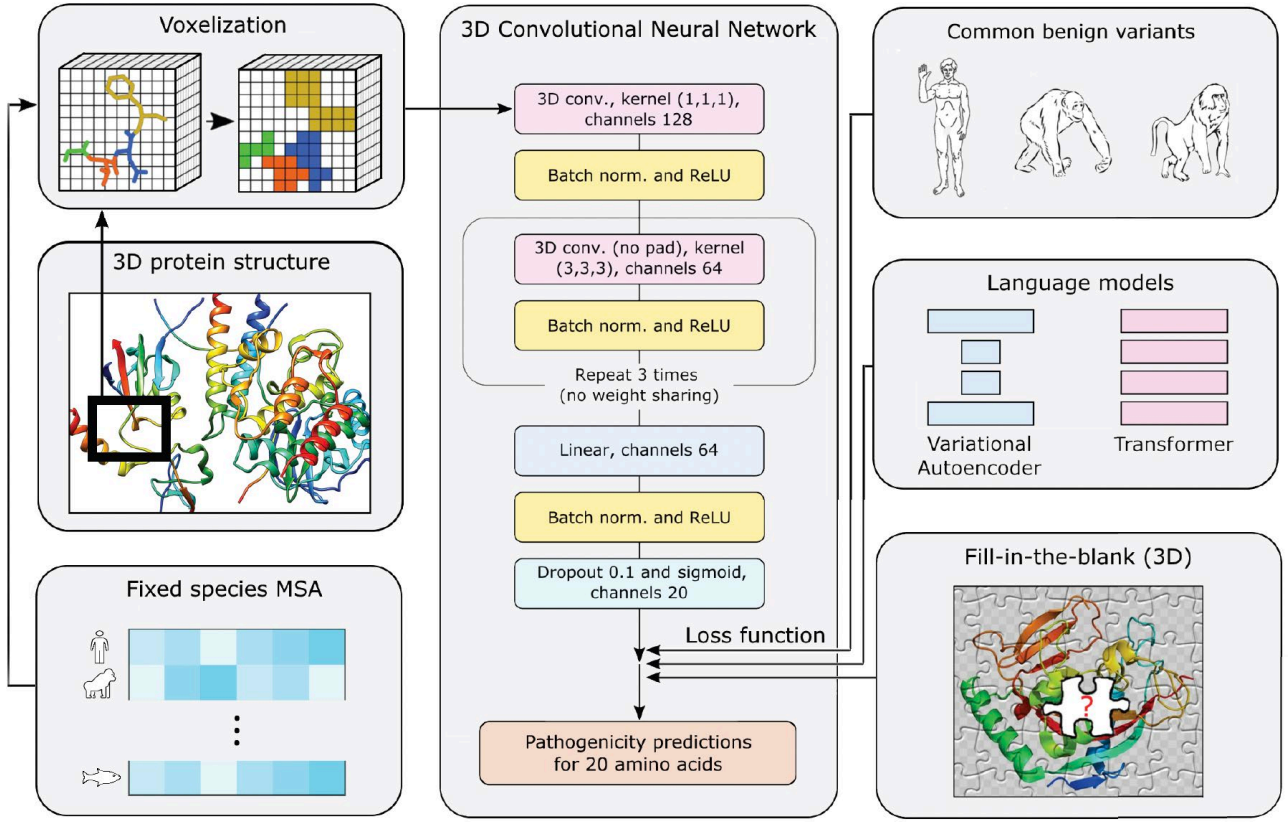
illumina®

Fiziev et al, Science 2023; Gao et al, Science 2023

PrimateAI-3D: state-of-the-art 3D-NN for variant interpretation



Benign and pathogenic variants localize to different regions of 3D protein structure

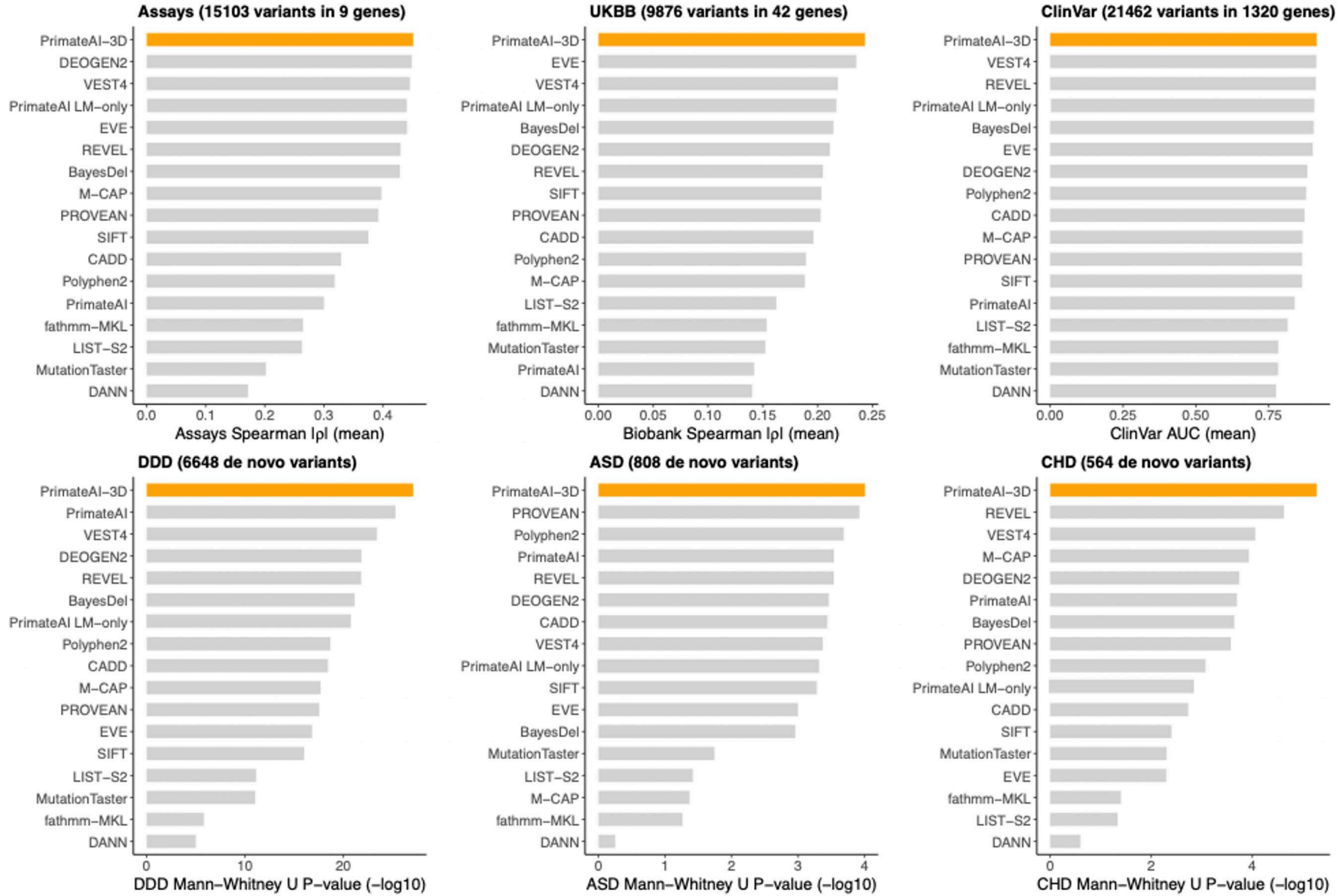


PrimateAI-3D is a 3-D convolutional neural network for missense variant pathogenicity prediction trained on 4.3 million protein-altering variants from across 234 primate species (including human)

illumina®

Fiziev et al, Science 2023; Gao et al, Science 2023

PrimateAI-3D leads in all clinical variant interpretation benchmarks

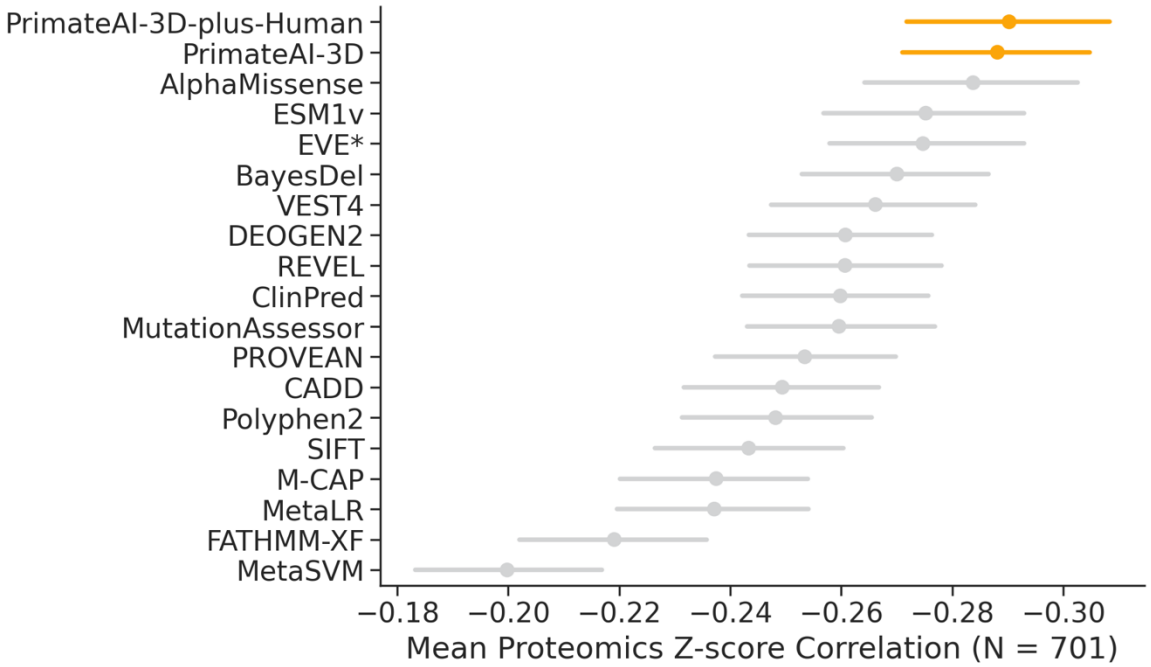


illumina®

Fiziev et al, Science 2023; Gao et al, Science 2023

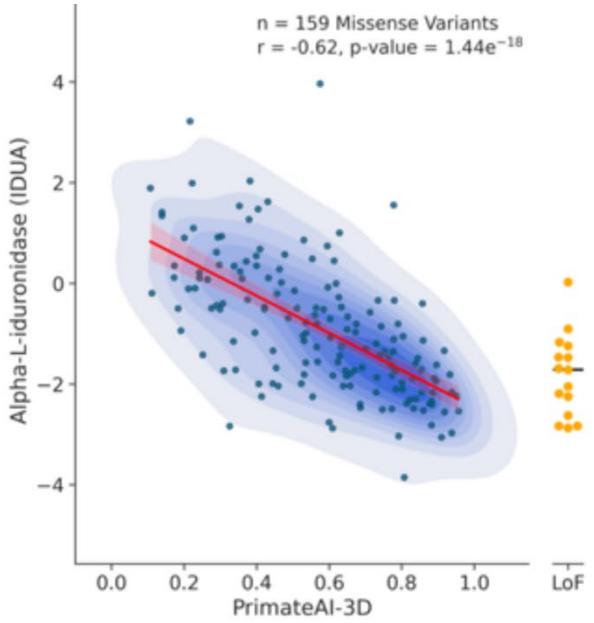
Using protein levels to benchmark AI variant prediction algorithms

n=701 O-link proteins in 50,000 UKBB participants

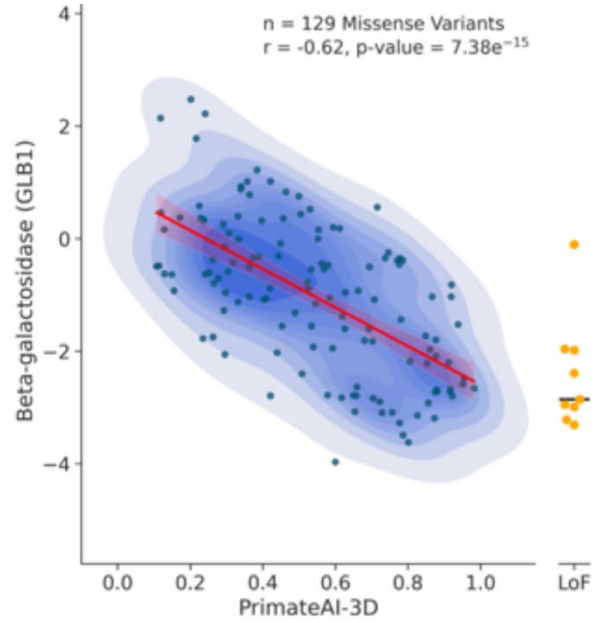


Correlation between in blood plasma protein levels and classifier score

IDUA



GLB1

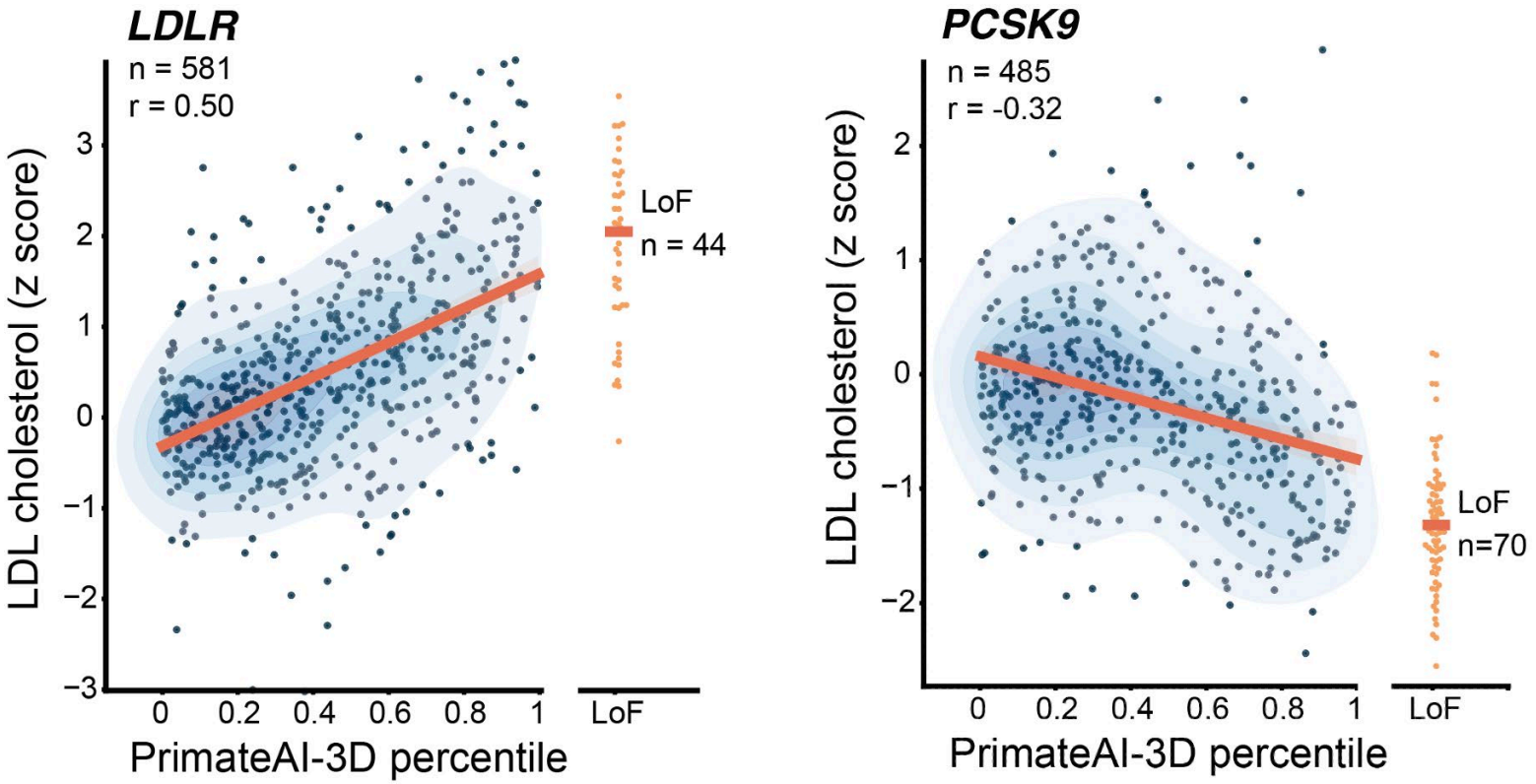


PrimateAI-3D scores for variants in disease-associated proteins IDUA and GLB1

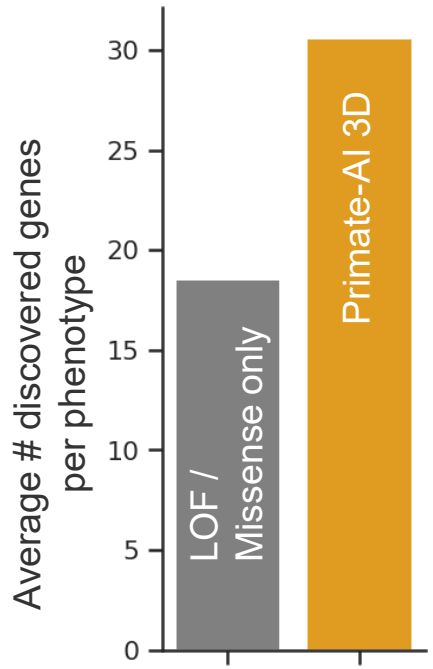
illumina®

Variant interpretation tools improve target discovery power

Clinical biomarkers highly correlate with PrimateAI-3D prediction scores

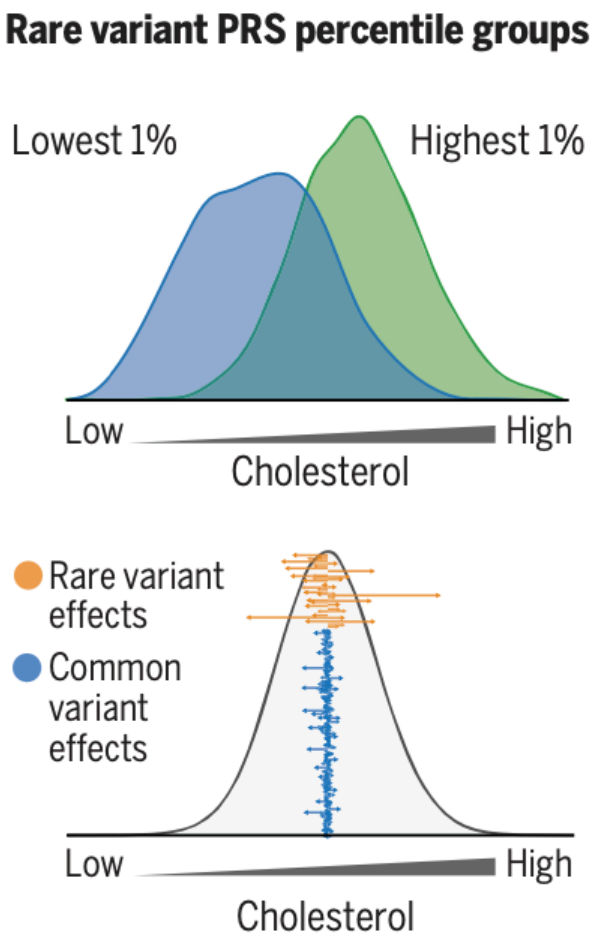


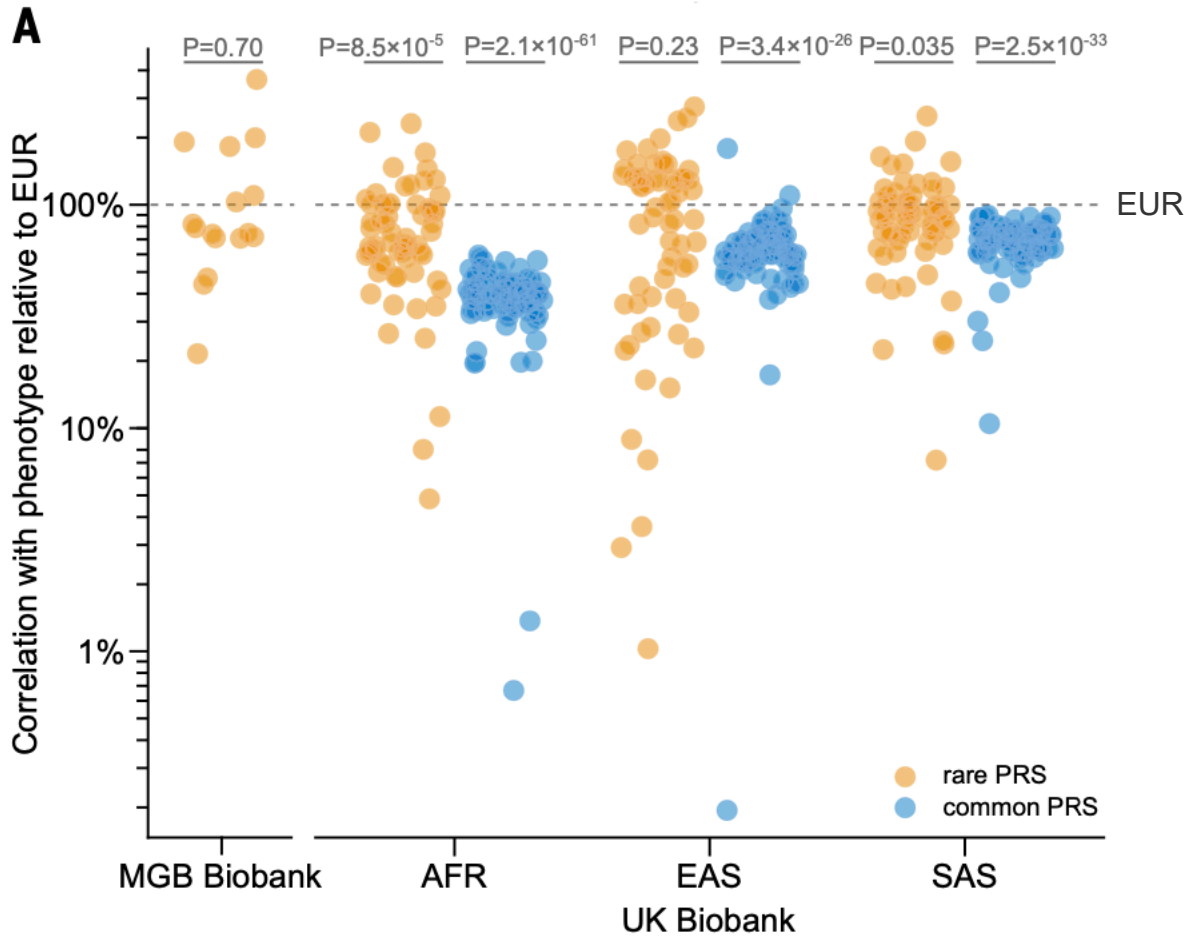
Using PrimateAI-3D predictions in rare variant burden test finds 64% more gene-phenotype associations



Fiziev et al, Science 2023; Gao et al, Science 2023

Rare variant PRS generalizes well to non-European populations





illumina®

Fiziev et al, Science 2023; Gao et al, Science 2023

Variant interpretation is a key bottleneck for drug target discovery

Develop drugs that mimic natural genetic variants

Power to discovery novel drug targets =

Variant Interpretation
X
Cohort Size

Identification of genetic variants that contribute to human disease
helps improve the odds of success for drug discovery and clinical trials

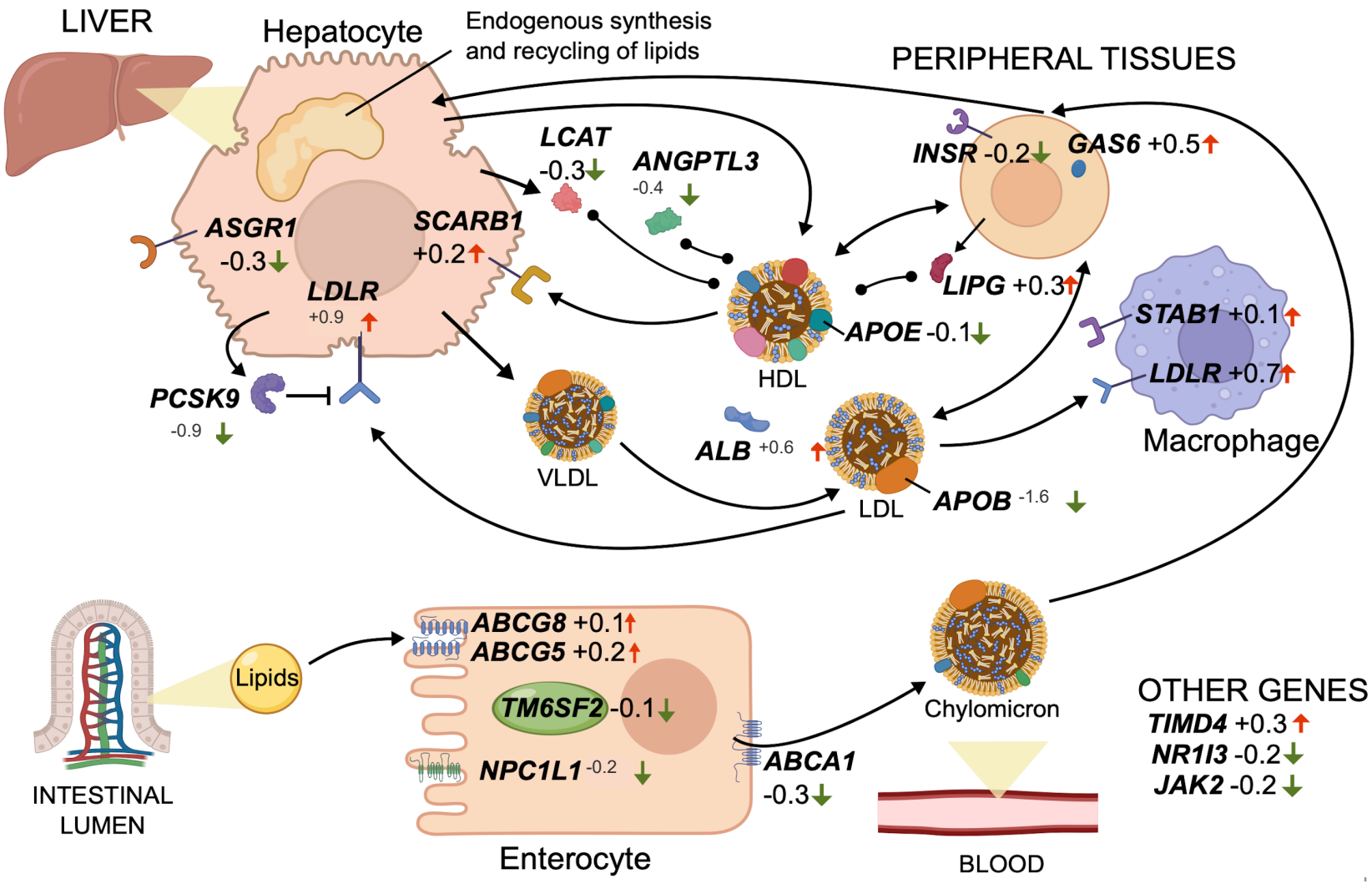
Recovery of cholesterol pathway from AI + genetics

Genes associated with Cholesterol in the UK Biobank 450k WES cohort

Rank	Gene	FDR	Effect	# Carriers
1	PCSK9	0.E+00	-0.9	1592
2	LDLR	0.E+00	0.9	1380
3	APOB	0.E+00	-1.6	401
4	ANGPTL3	8.E-90	-0.4	1644
5	NPC1L1	6.E-47	-0.2	4606
6	ABCG5	4.E-39	0.2	4479
7	ABCA1	3.E-30	-0.2	4105
8	ASGR1	7.E-28	-0.3	1311
	…			

48 genes with exome-wide significance

Loss of function results in
- Decreased cholesterol
- Increased cholesterol



Fiziev et al, Science 2023; Gao et al, Science 2023

PrimateAI-3D recovers all cholesterol drugs using UKBB 450K exomes

PCSK9 inhibitors ($1B market)

ANGPTL3/4 inhibitors

NPC1L1 inhibitors (>$1B market)

Rank	Gene	FDR	Effect	Druggable	# Carriers	
1	PCSK9	0.E+00	-0.9	TRUE	1592	
2	LDLR	0.E+00	0.9	TRUE	1380	
3	APOB	0.E+00	-1.6	TRUE	401	
4	ANGPTL3	8.E-90	-0.4	TRUE	1644	
5	NPC1L1	6.E-47	-0.2	TRUE	4606	
6	ABCG5	4.E-39	0.2	FALSE	4479	
7	ABCA1	3.E-30	-0.2	TRUE	4105	
8	ASGR1	7.E-28	-0.3	TRUE	1311	
9	ALB	2.E-18	0.6	TRUE	205	
10	ABCA6	2.E-16	0.1	FALSE	4487	
11	TIMD4	4.E-14	0.3	FALSE	828	
12	TM6SF2	8.E-13	-0.1	FALSE	3744	
13	RRBP1	2.E-09	-0.3	FALSE	381	
14	DENND4C	6.E-09	0.2	FALSE	1217	
15	CETP	3.E-08	-0.1	TRUE	2386	
16	STAB1	1.E-07	0.1	FALSE	5014	
17	ABCG8	3.E-07	0.1	FALSE	4566	
18	HBB	7.E-07	-0.4	TRUE	211	Anemia
19	MYLIP	8.E-06	-0.2	FALSE	775	
20	NR1H4	1.E-05	-0.2	TRUE	582	
21	PDE3B	3.E-05	-0.1	TRUE	2594	
22	GAS6	4.E-05	0.2	TRUE	470	
23	DENND4A	4.E-05	0.2	FALSE	974	
24	GPAM	5.E-05	-0.2	FALSE	841	

Rank	Gene	FDR	Effect	Druggable	# Carriers
25	HMGCR	7.E-05	-0.1	TRUE	1441
26	APOE	7.E-05	-0.1	TRUE	2170
27	KEAP1	7.E-04	0.2	TRUE	844
28	B4GALT1	9.E-04	-0.4	TRUE	130
29	CD36	1.E-03	-0.1	TRUE	3625
30	INSR	2.E-03	-0.1	TRUE	1982
31	G6PC	2.E-03	0.1	TRUE	1901
32	RBM47	3.E-03	0.2	FALSE	365
33	SP1	3.E-03	-0.2	FALSE	550
34	ANLN	4.E-03	-0.2	FALSE	436
35	PLA2G12B	5.E-03	-0.4	TRUE	113
36	SLC4A1	5.E-03	-0.3	TRUE	265
37	APOA5	1.E-02	0.2	TRUE	517
38	FOXA3	1.E-02	-0.1	FALSE	1033
39	SBNO2	2.E-02	-0.05	FALSE	7342
40	SYNJ2BP	3.E-02	-0.2	FALSE	283
41	PDSS2	3.E-02	-0.3	FALSE	259
42	FCGRT	3.E-02	0.2	TRUE	658
43	ANGPTL4	3.E-02	-0.1	TRUE	1413
44	SPHK1	3.E-02	-0.1	TRUE	1686
45	TBC1D8	3.E-02	-0.1	FALSE	2380
46	LIPG	3.E-02	0.2	TRUE	426
47	CPT1A	5.E-02	-0.1	TRUE	1119
48	SHMT1	5.E-02	-0.1	FALSE	1756

Statins pathway ($14B market)

Loss of function results in
- Decreased cholesterol
- Increased cholesterol

illumina

Fiziev et al, Science 2023; Gao et al, Science 2023

AI algorithms in development for understanding the noncoding genome

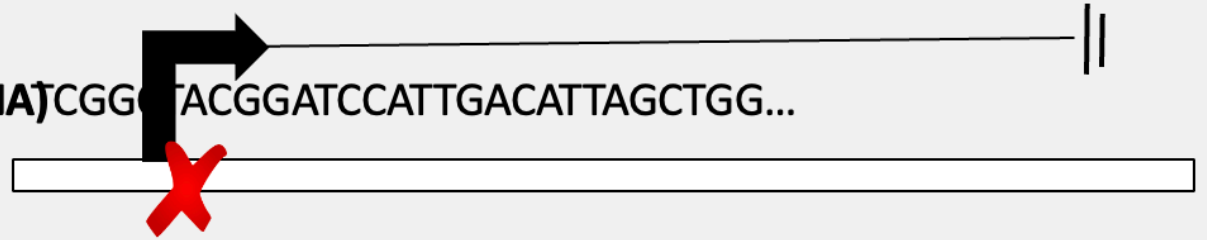
Reduce VUS in coding regions

Translated protein

PrimateAI-3D
Errors in protein function

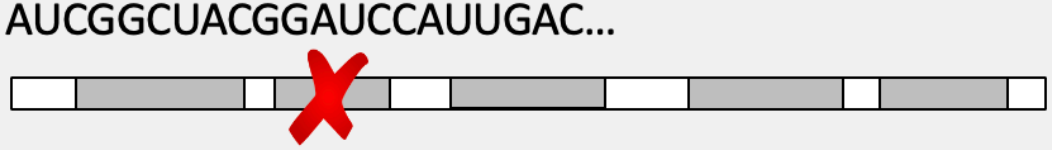
Non-coding interpretation

Transcription (DNA)



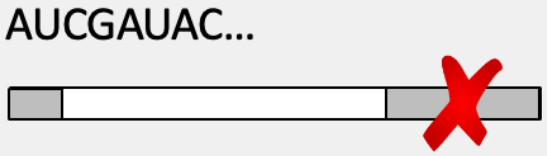
PromoterAI / methylAI
Gene not expressed

Splicing (RNA)



SpliceAI v2
Errors in specifying the protein-coding sequence

RNA decay (RNA)

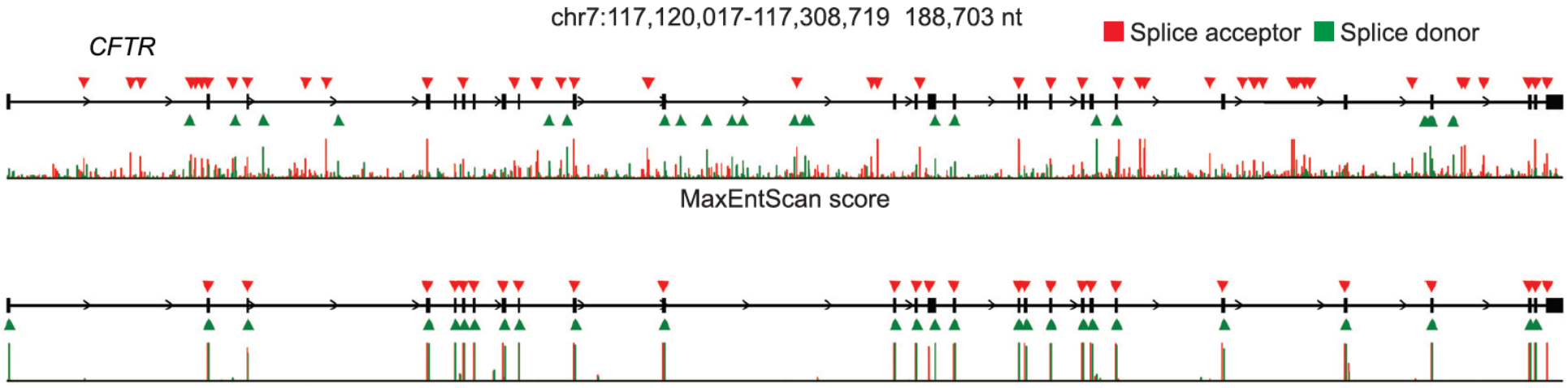


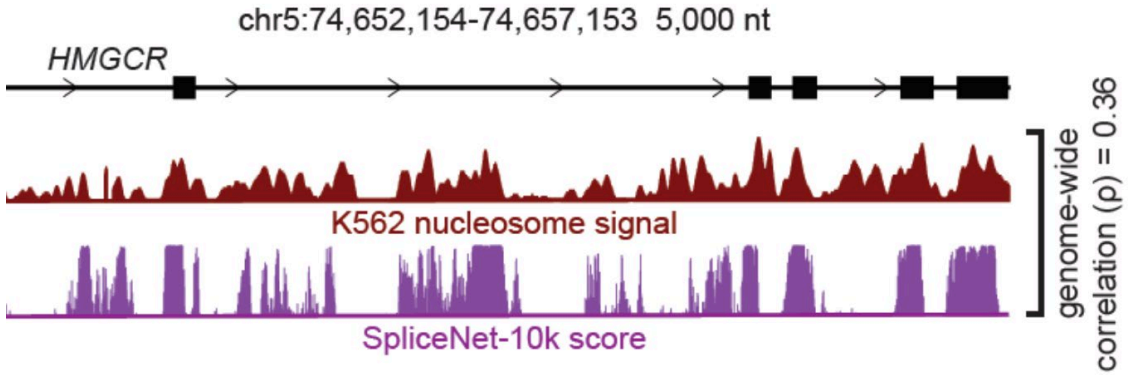
UTRs / microRNAs
Altered stability & translation

SpliceAI predicts noncoding variants that disrupt splicing

Best previous algorithm *(**18%** accuracy)*

SpliceAI deep neural network *(**95%** accuracy)*



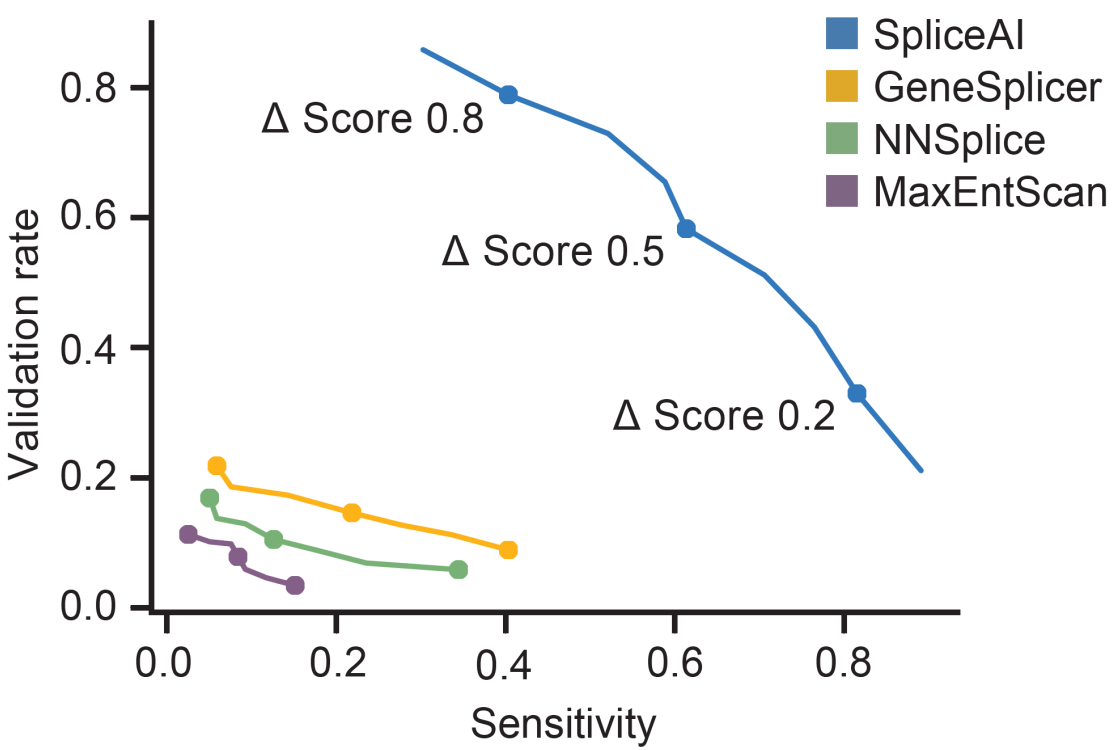


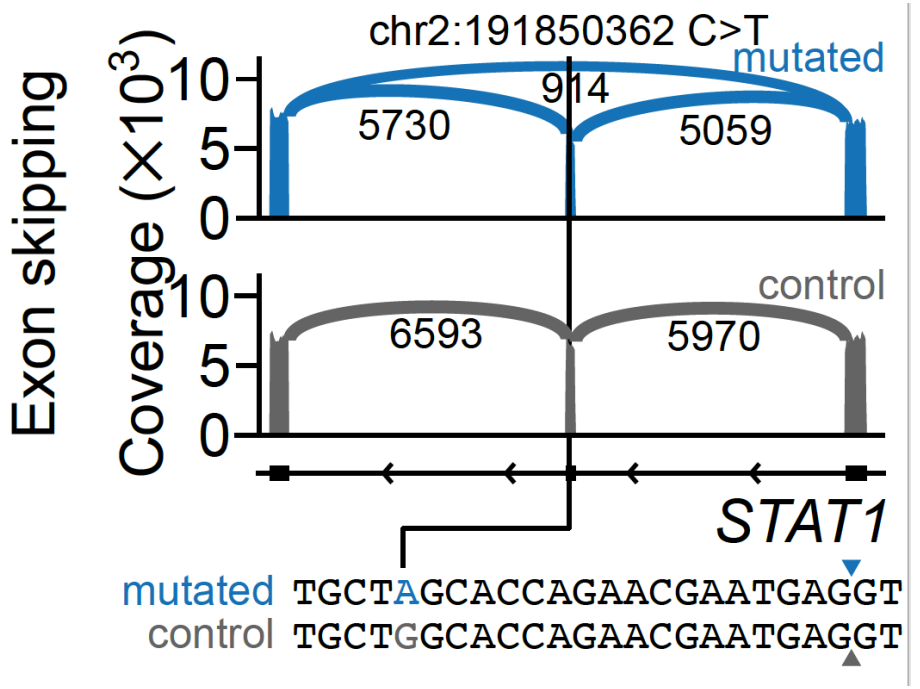
- Long range determinants up to 10kb are crucial for splicing specificity
- Intron / exon length, nucleosome positioning play major roles

illumina®

Jaganathan et al, Cell 2019

Clinical validation of SpliceAI noncoding pathogenic variants





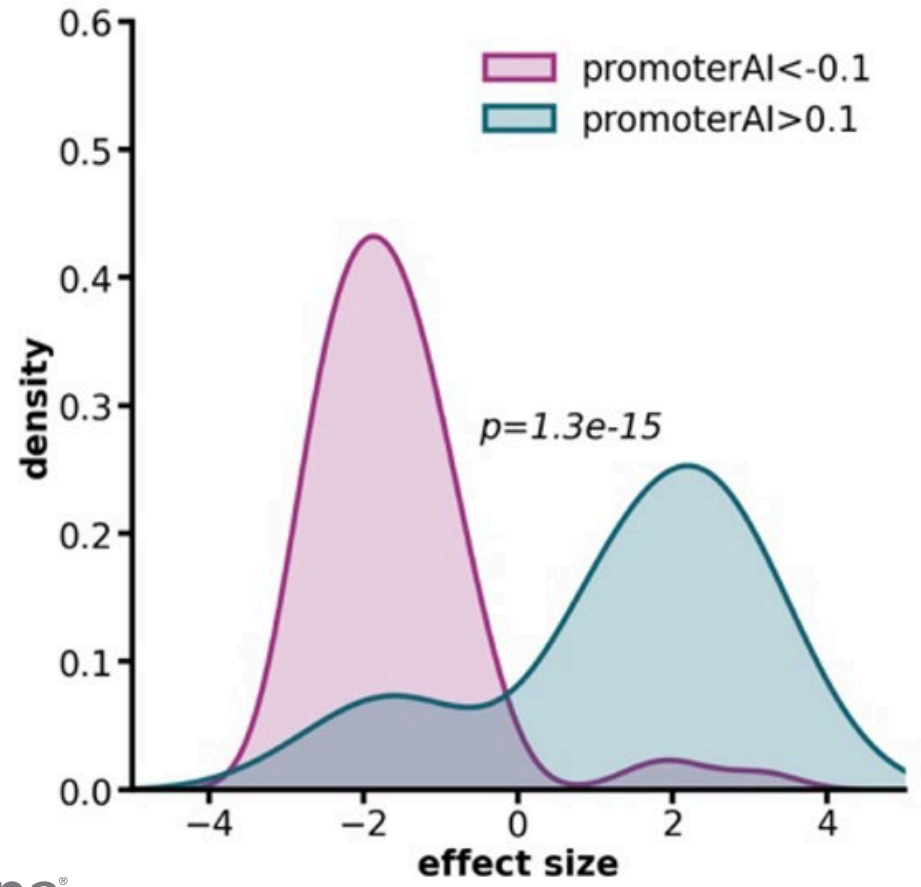
- RNA-seq Validation Experiments with Stephan Sanders (UCSF) –
  - Used deep learning to predict noncoding mutations in 28 undiagnosed autism patients
  - RNA-seq in patient blood samples validated the predicted aberrant splice event in 75% of case
  - Similar validation rate in Genomics England on 5000 undiagnosed rare disease patients

illumina®

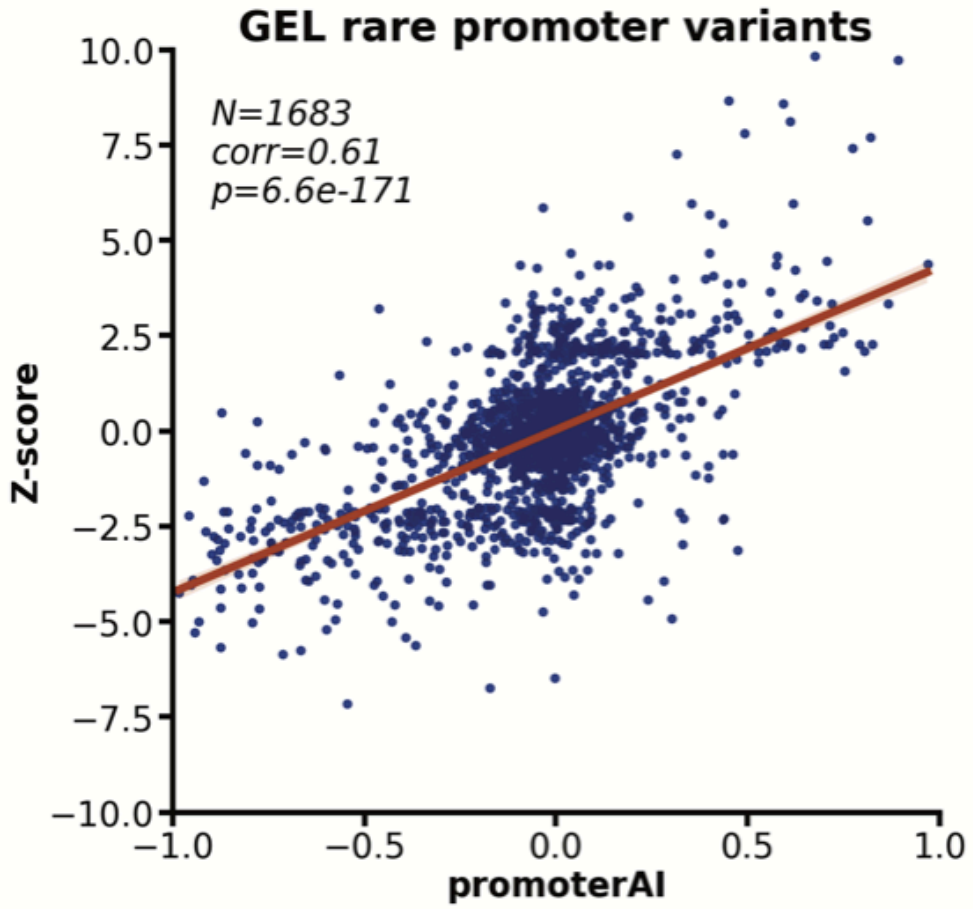
Jaganathan et al, Cell 2019

PromoterAI variants produce outlier gene expression on both the RNA and protein levels

Validation in 50,000 individuals with protein-omics data in UKBB



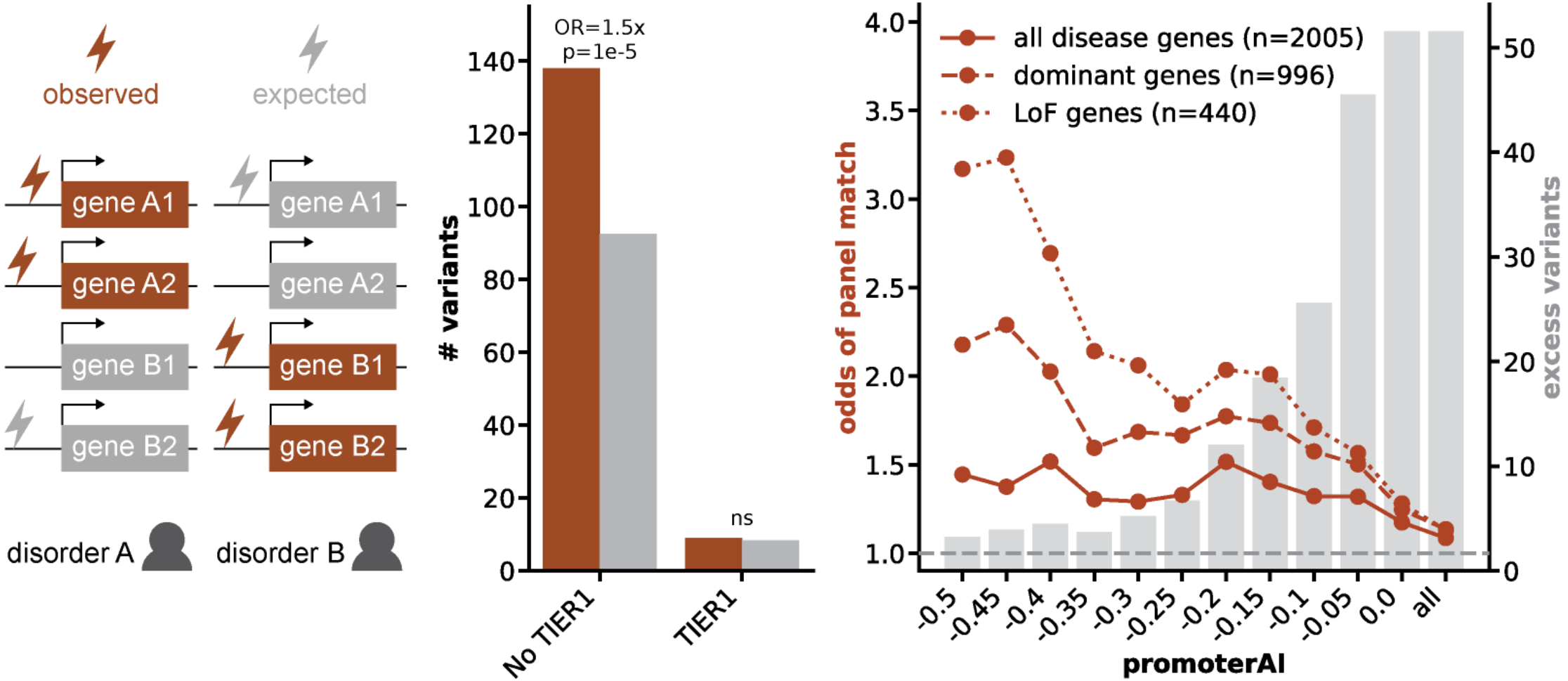
Validation in 5,000 individuals with RNA-seq in Genomics England



illumina®

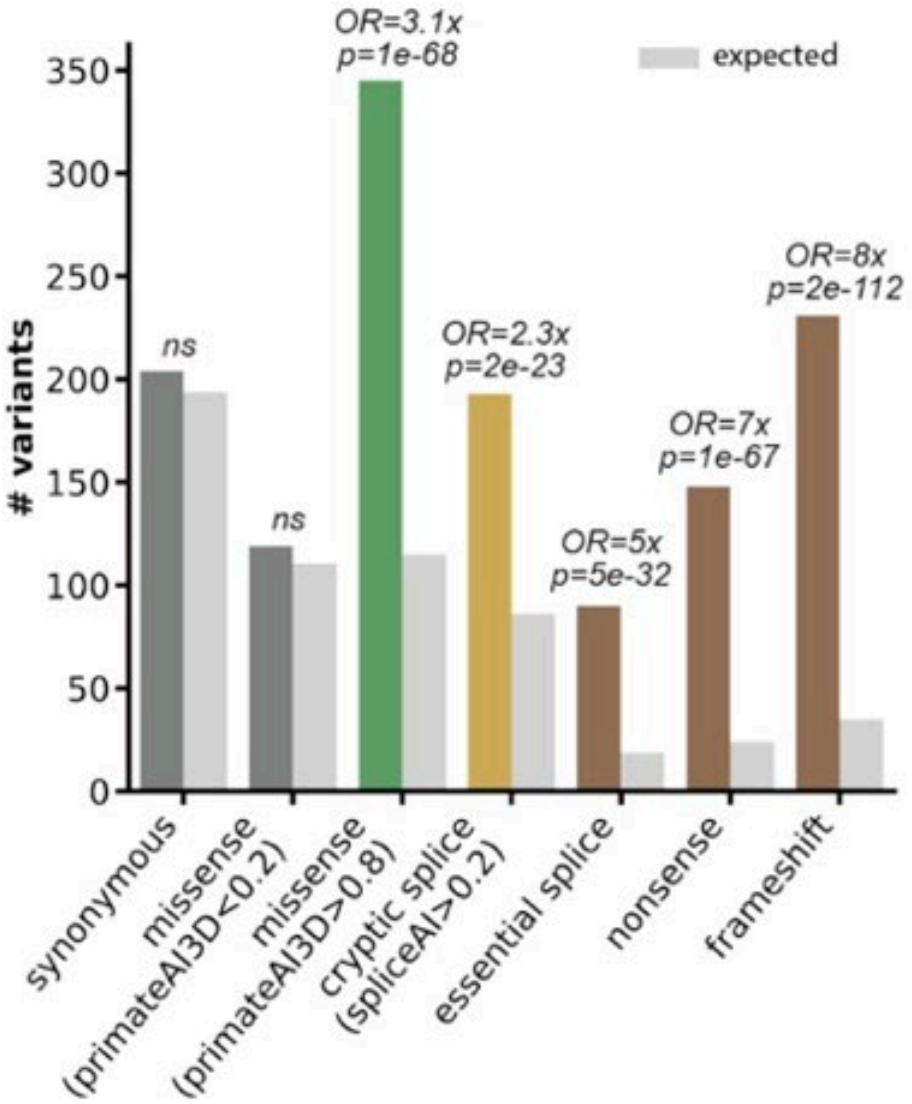
Jaganathan et al, Science 2025

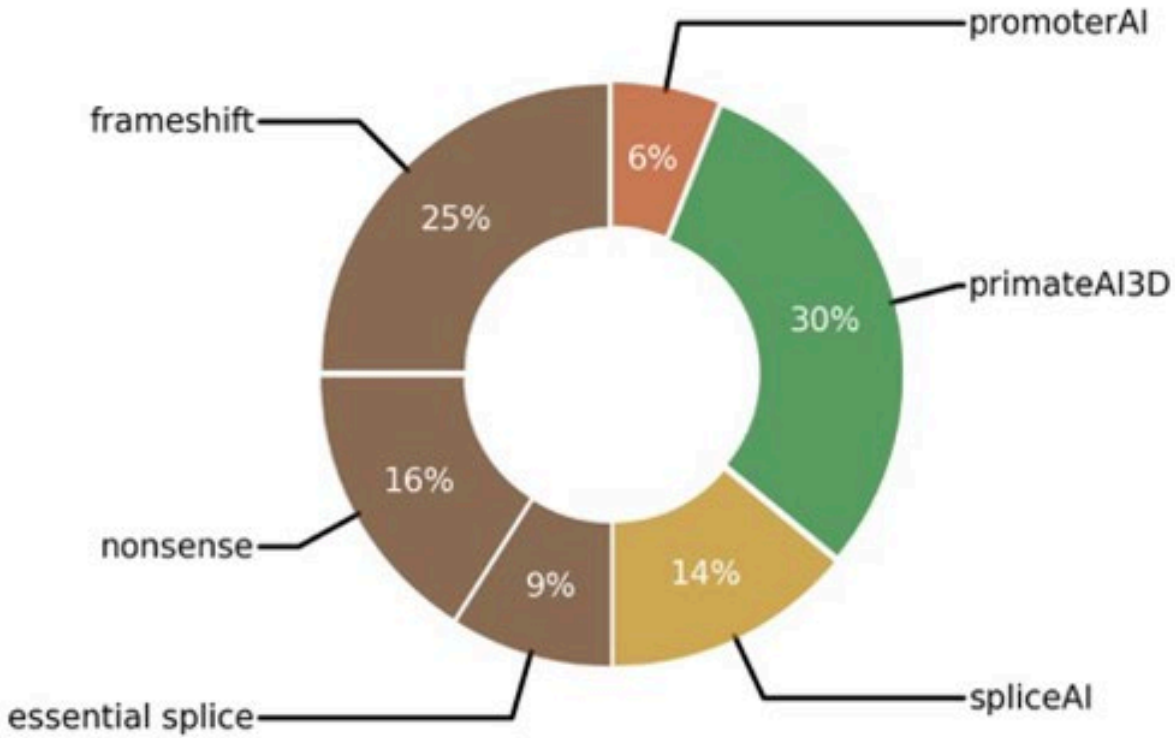
PromoterAI variants are enriched in disease genes in the 100,000 Genomics England rare disease cohort



Jaganathan et al, Science 2025

Contribution of coding & noncoding SNPs to GEL rare disease





illumina®

Jaganathan et al, Science 2025

100,000,000+ cell Perturb-seq atlas by end of 2025

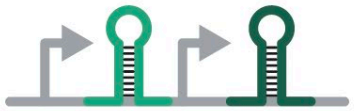
Selected cancer cell lines covering diverse cell types

HAP1	#HT29
THP-1	HCT116
A549	NCI H460
HEK293T	K562
*HepG2	HeLa
H4	

IPSC-derived cell lines, embryoid bodies, organoids

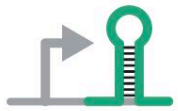
Primary immune cells

Dual Guide RNA



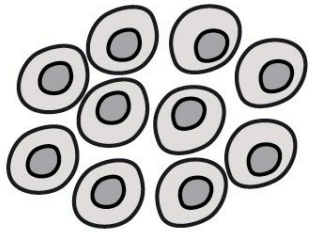
or

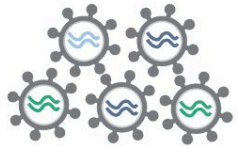
Single Guide RNA



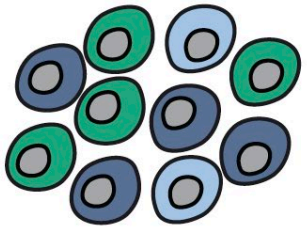
Systematically optimize time course, guides-per-cell, guide sequence design

Cells expressing CRISPRi/a machinery



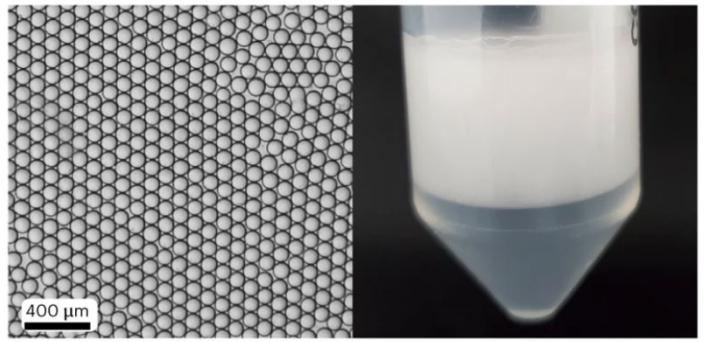


Genome-wide CRISPRi/a Perturb-seq



1M cell Fluent PIP-seq kit with direct guide capture

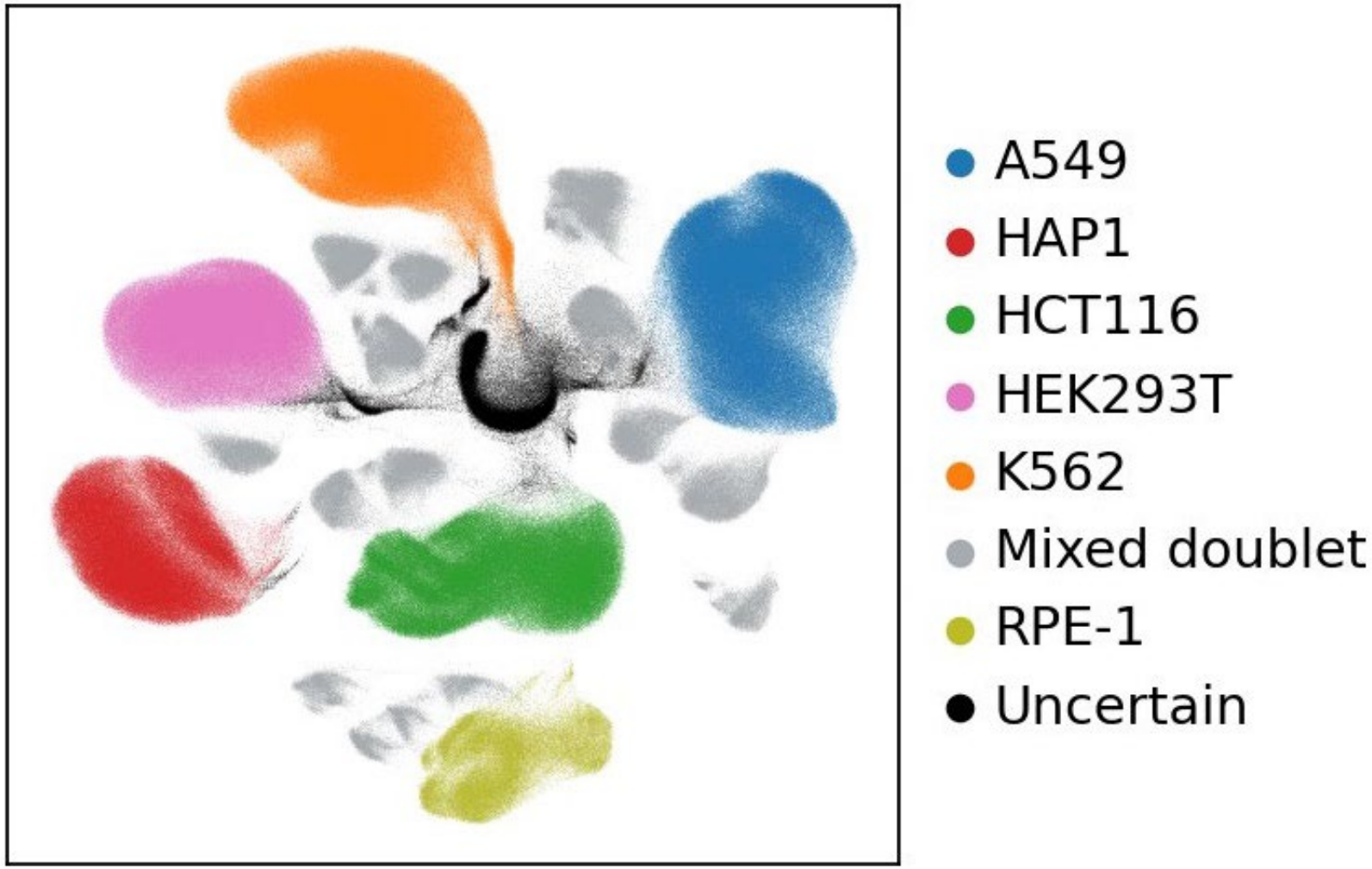
50-ml Falcon (10 ml PIPs ~1 million cells cell input)

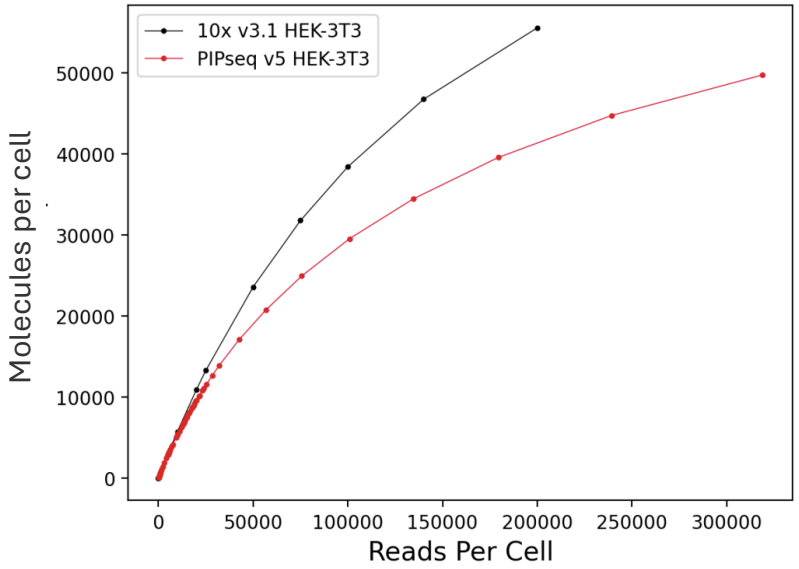


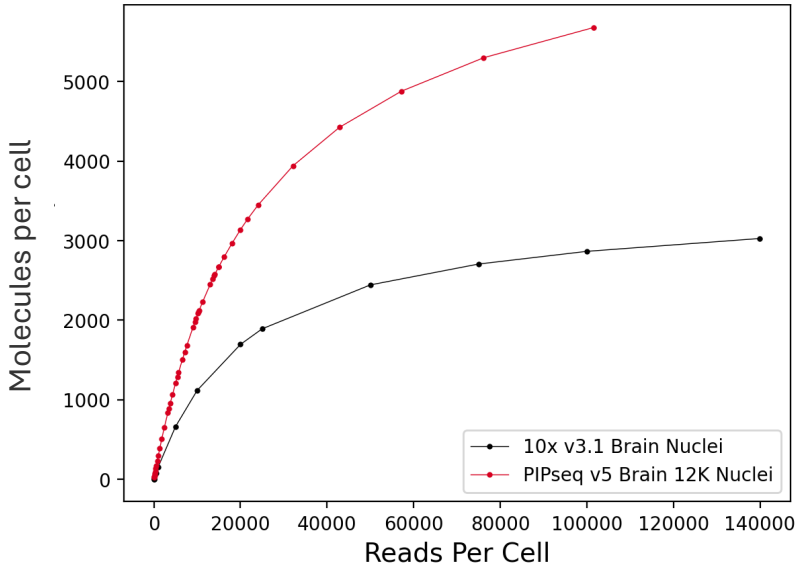
Overloading the 1M kit to get 1.7M single cells in a single experiment

Same amount of data would require ~100 overloaded 10X runs

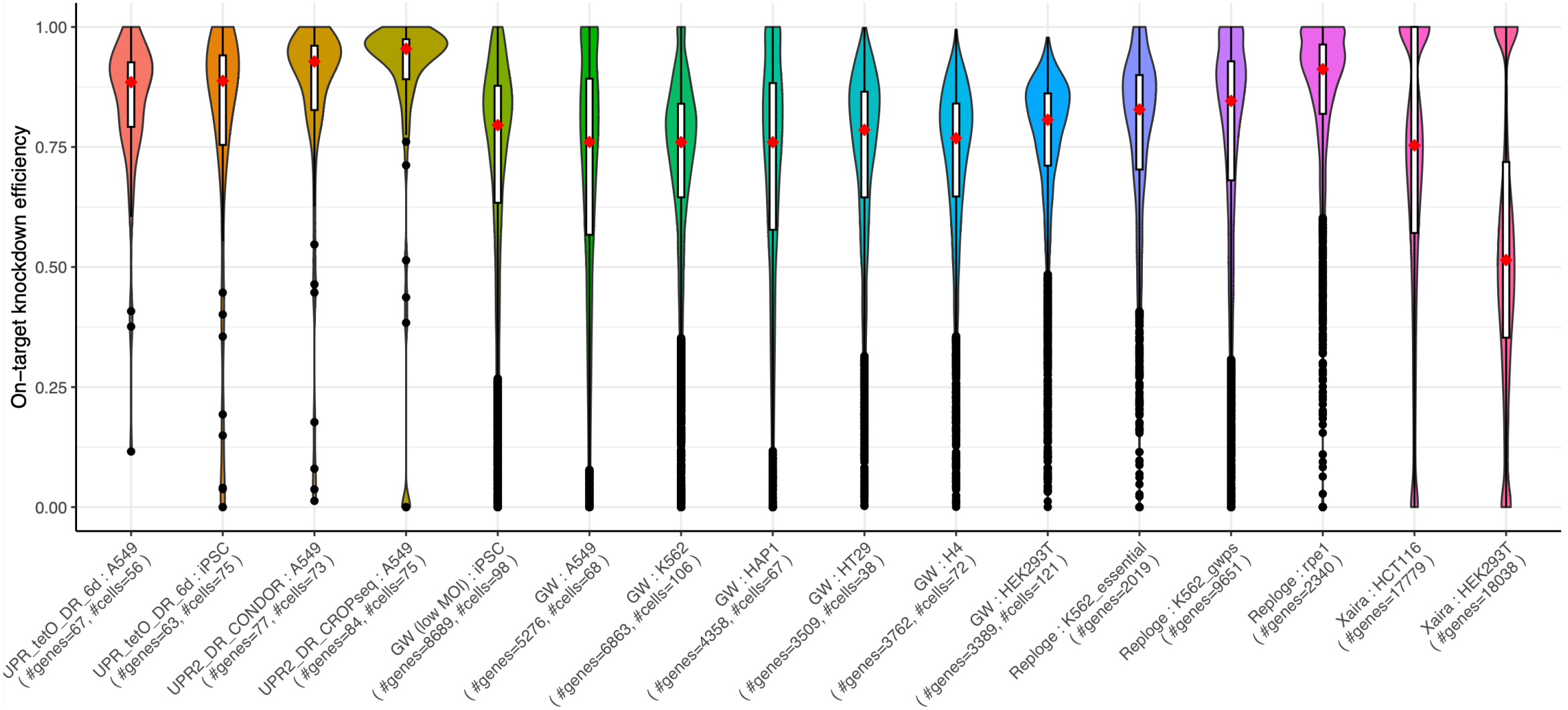
4M cells loaded – 1.7M recovered







CRISPRi perturb-seq knockdown vs previously published data



illumina®

Illumina UPR pilots (90% KD)

Illumina Genome-Wide (80% KD)

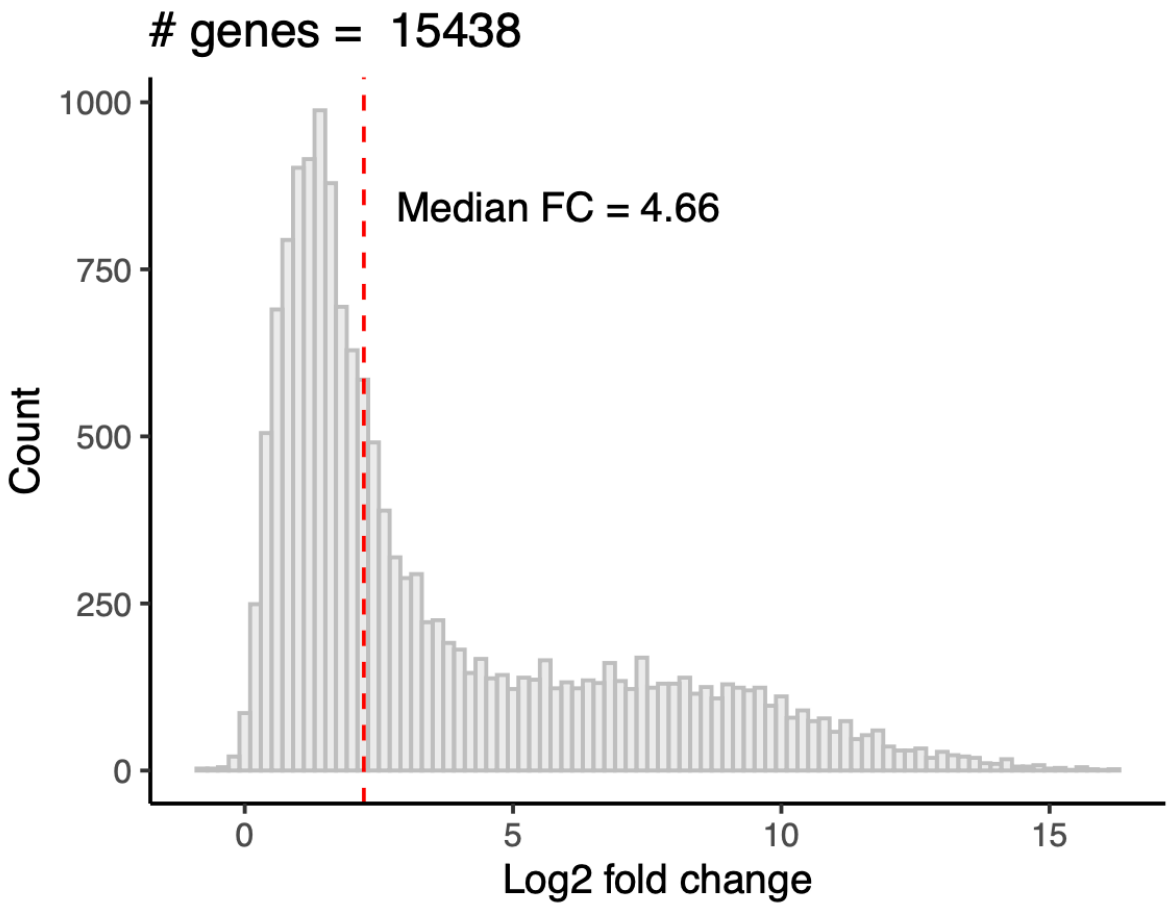
Replogle et al (80-90% KD)

Xaira (50-75% KD)

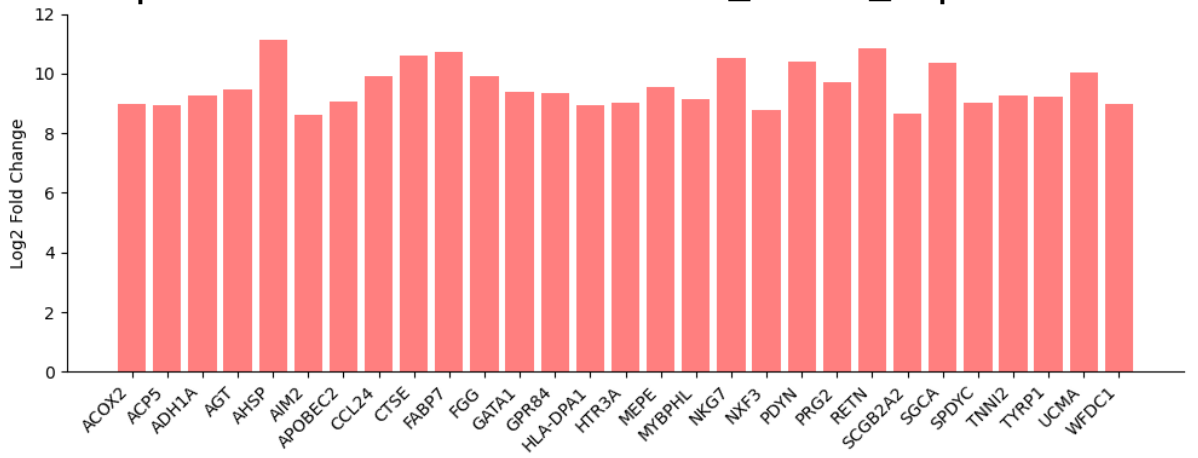
Genome-wide CRISPRa perturb-seq targets 6M cells in HAP1 with 60,000 guide pairs

3-fold genome-wide for library optimization

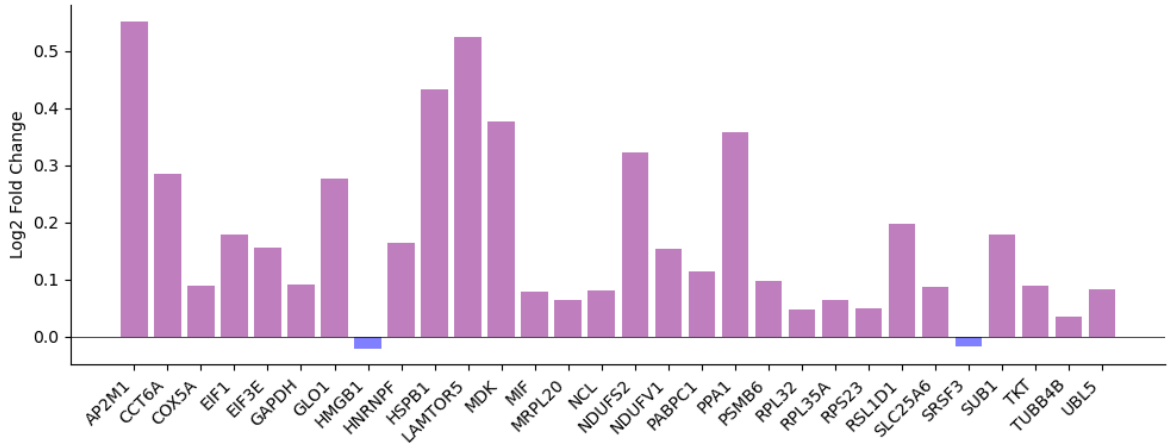
Showing 15438 genes with minimal amount of expression in either perturbed or control cells (> 0.1)



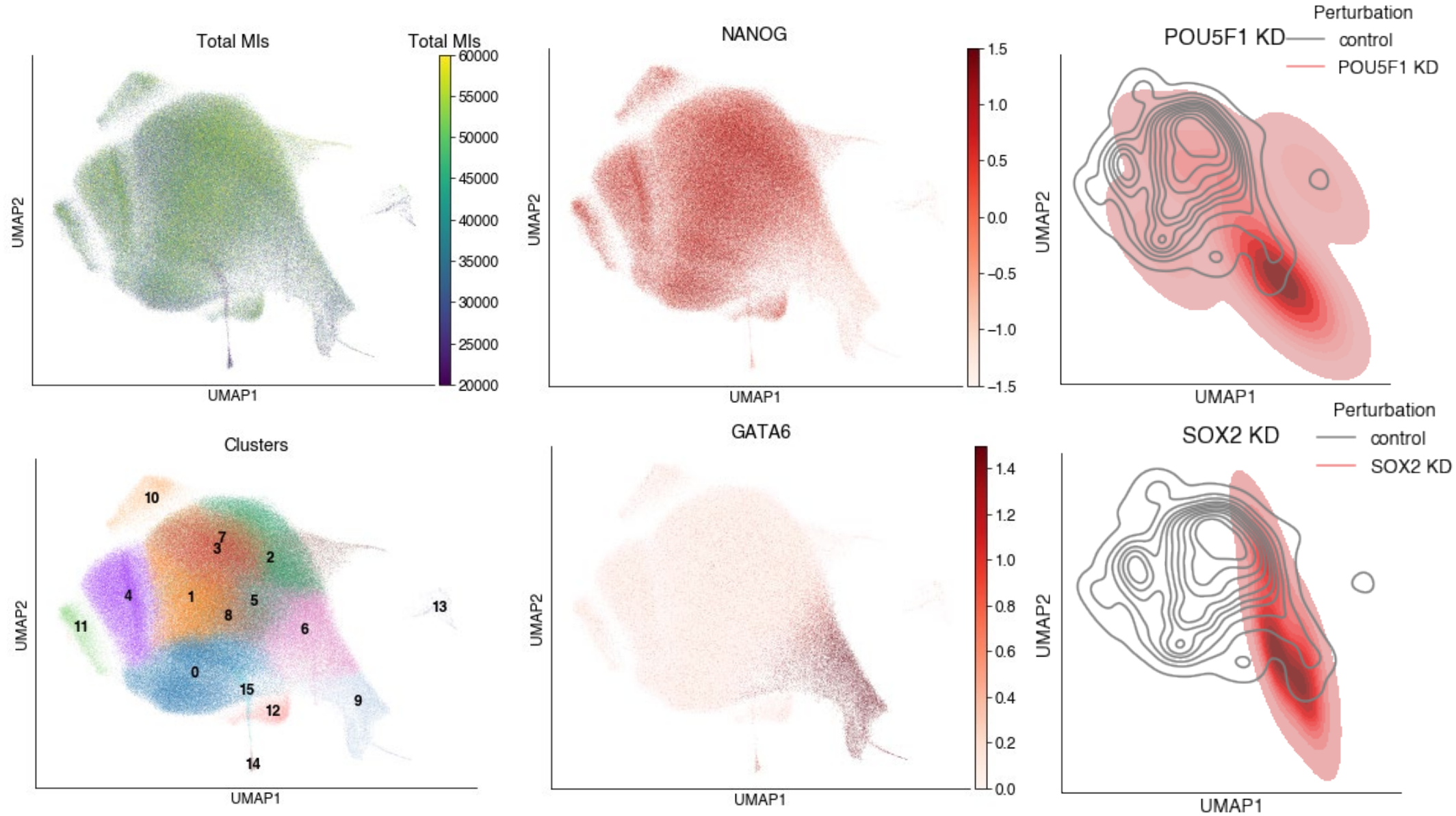
Top 30 most highly activated genes with #perturbed cells >100 and ctrl_mean_expr < 0.001



Top 30 highly expressed genes in controls with #perturbed cells >100



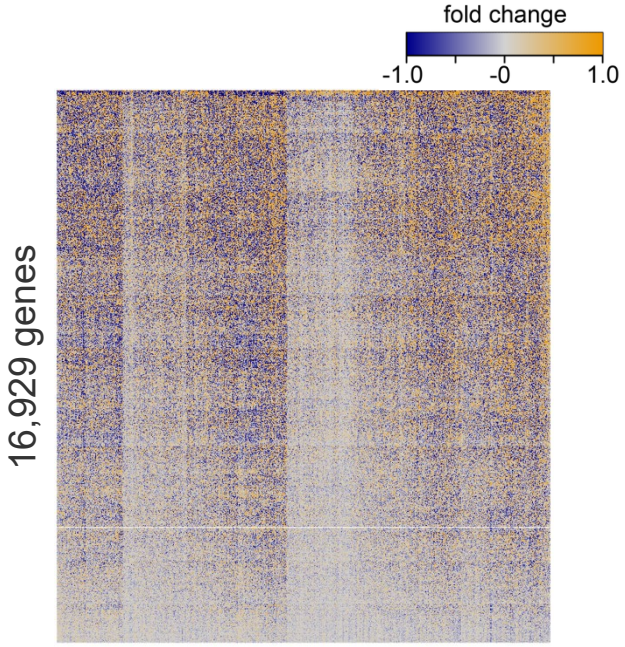
IPSC perturb-seq: Knockdown of OCT4, SOX2 lead to differentiation



illumina®

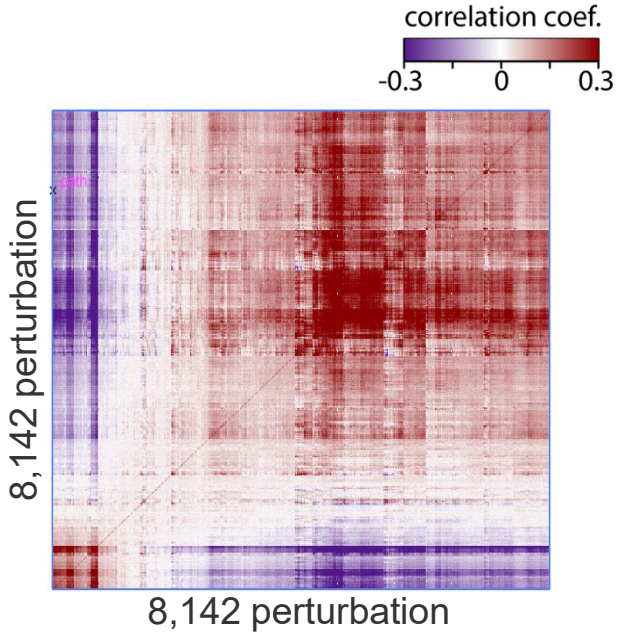
Perturbation correlation identifies genes with similar functions (demonstrated by protein-protein interactions)

Expression profile upon gene perturbation



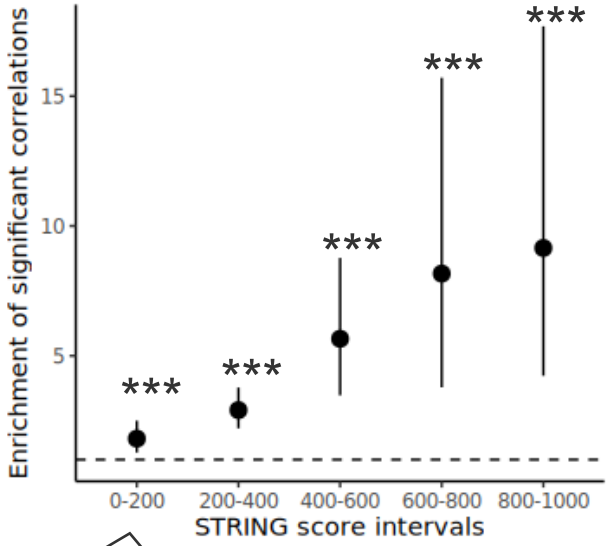
pairwise correlation of expression profiles

Perturbation-perturbation correlation matrix



gene pairs with similar functions (Bonferroni p <0.05 & coef. >0.1)

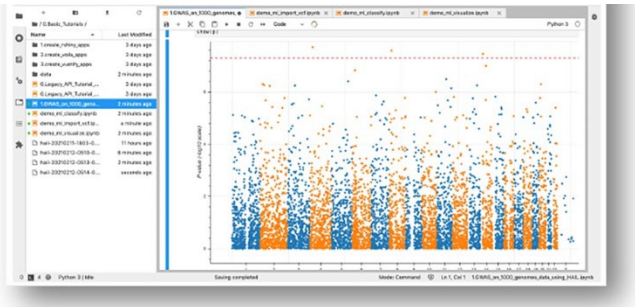
Significant correlations enriched in STRING PPIs

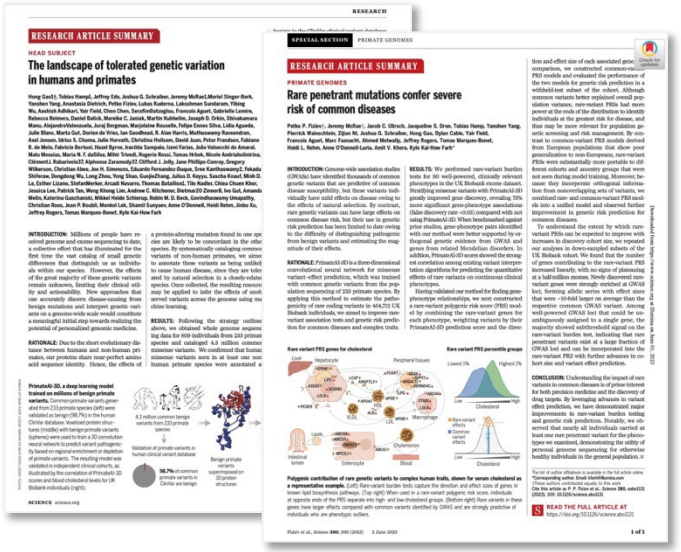


illumina®

Illumina's solutions for precision medicine and discovery

DRAGEN





1 Sequencing platform

- Instrument
- Library prep
- Reagents

2 Software ecosystem

- DRAGEN™ secondary analysis
- Illumina Connected Analytics (ICA)
- Emedgene™
- Illumina Connected Insights (ICI)
- Illumina Connected Multiomics (ICM)

3 AI & perturb-seq

Leading AI algorithms and datasets for delivering new insights into human genetics at scale