Functional interpretation of Genome-wide association studies

(GWAS2Function)

Kushal K. Dey

Background to Human Genetic studies

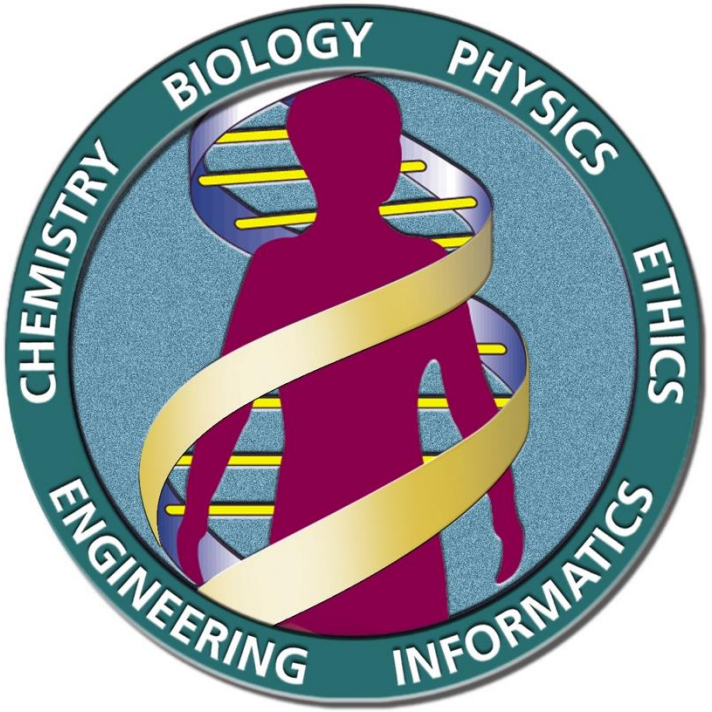
The Human Genome Project

- Human Genome Project (launched 1990, completed 2003)

- Generate the first sequence of the human genome

    - *Reference genome*: all base pairs in human genome

    - *Map all genes* – observed ~22K protein-coding genes



Got the ball rolling in terms of genomic sequencing

HapMap Project: Cataloguing variations in the sequences of human DNA (2002-2010) (1,000 individuals)

DNA sequence of any two individuals is 99.5% similar, however the 0.5% difference drives differences in physiological traits and disease risk.

HapMap catalogued variation across ~1,000 individuals.

Sites in the DNA sequence where individuals differ at a single DNA base are called **single nucleotide polymorphisms (SNPs).**

SNPs were identified at specific chromosomal positions (what nomenclature to use?)

	chrom.	physical position (bp)
rs10910034	1	2165898
rs1713712	1	2166021

Genome wide Association studies

Collecting genotype and phenotype data from many many individuals (order of 100, 000 individuals)

Large-scale retrospective studies (~100K-1M individuals)

Collecting genotype and phenotype data from many many individuals (order of 100, 000 individuals)

Large-scale retrospective studies (~100K-1M individuals)

biobank uk
Improving the health of future generations

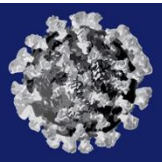
BioBank Japan
バイオバンク・ジャパン

All of Us℠ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

MVP 10 YEARS
MILLION VETERAN PROGRAM

Disease-related prospective studies (~10K-100K)

PGC
Psychiatric Genomics Consortium


The COVID-19
Host Genetics Initiative

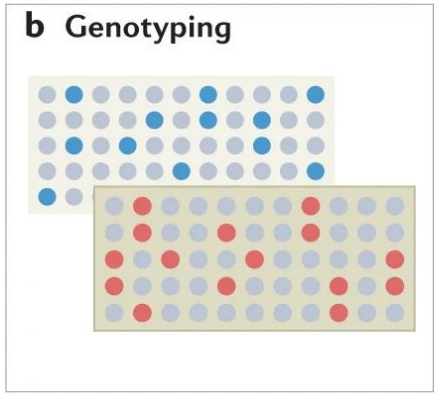
CARDIoGRAMplusC4D

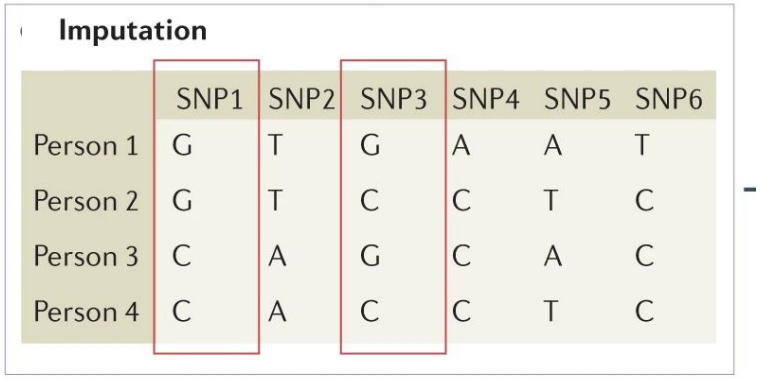
Background: Genome-wide Association Studies

biobank uk
Improving the health of future generations

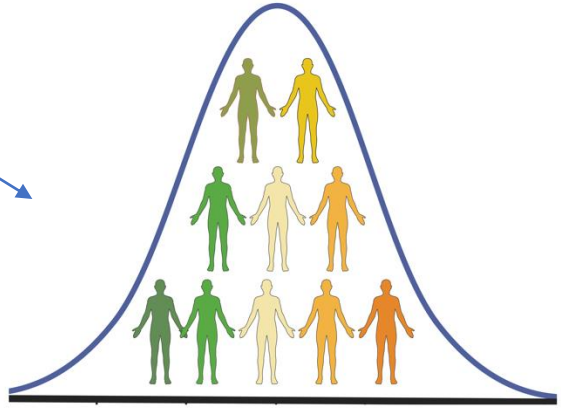
All of Us The Precision Medicine Initiative
THE FUTURE OF HEALTH BEGINS WITH YOU

PGC
Psychiatric Genomics Consortium

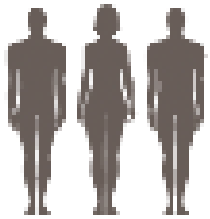


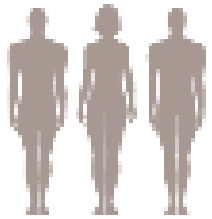


	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C



Quantitative phenotype (red blood cell count, LDL cholesterol)

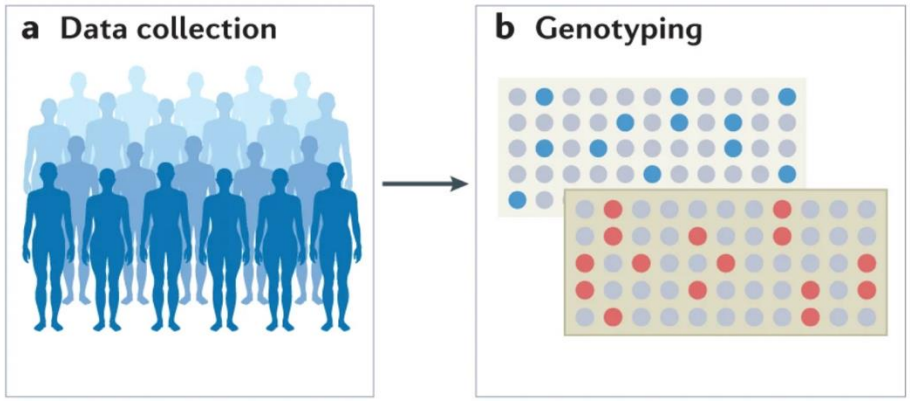




Cases Controls

Alzheimers, Schizophrenia, Cancers

Mathematical model for Genome Wide association studies



Sequencing strategies:
SNP array + imputation
Whole exome sequencing and
Whole Genome sequencing

Phenotype

Global ancestry changes

$$\boldsymbol{Y} \sim \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + g + e$$

(linear or logistic regression)

Genotype effect

$$g \sim N(0, \sigma^2_{\mathrm{A}}\boldsymbol{\psi})$$

Local population relatedness

$$e \sim N(0, \sigma^2_e\boldsymbol{I})$$

Y vector of phenotype values for all N individuals (for example: height or 1/0 for Type 2 diabetes status)

X_s vector of genotype values for all N individuals at SNP s (0/1/2 for unscaled: ore standardized)

W matrix of covariates (age, sex, ancestry PCs)

g – represents polygenic effect of other SNPs
e - random effect of residual errors
$\boldsymbol{\psi}$ kinship or genetic relatedness matrix

Calculating statistics from Genome Wide association studies

Estimates of the effect size

$$\hat{\beta}_{\text{snp}} = \frac{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{y}}{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{x}_{\text{snp}}} \text{with var} \left(\hat{\beta}_{\text{snp}} \right) = \frac{1}{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{x}_{\text{snp}}}$$

$$V = \sigma_g^2 \psi + \sigma_e^2 I$$

Overall phenotypic variance-covariance matrix = genetic + error

Obtain z scores and p-values of the effect based on this.

GCTA
a tool for Genome-wide Complex Trait Analysis

GCTA | SMR | GSMR | OSCA | CTG forum | Yang Lab

- Overview
- Download
- FAQ
- Basic Options
- GREML
- GWAS Analysis
  - MLMA
  - fastGWA

fastGWA

fastGWA: A fast MLM-based Genome-Wide Association tool

fastGWA is an ultra-efficient tool for mixed linear model (MLM)-based GWAS analysis of biobank-scale data such as the UK Biobank (see Jiang et al. *Nature Genetics* 2019 for details of the method). Credits: Longda Jiang (method, simulation and analysis), Zhili Zheng (method, software and analysis) and Jian Yang (method and overseeing).

We have applied fastGWA to 2,173 traits on 456,422 array-genotyped and imputed individuals and 2,048 traits on 49,960 whole-exome-sequenced (WES) individuals in the UK Biobank. All the summary statistics are available

Jiang et al 2019 *Nat Genet*

Calculating statistics from Genome Wide association studies

Estimates of the effect size

$$\hat{\beta}_{\text{snp}} = \frac{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{y}}{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{x}_{\text{snp}}} \text{with var} \left(\hat{\beta}_{\text{snp}} \right) = \frac{1}{\mathbf{x}_{\text{snp}}^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{x}_{\text{snp}}}$$

$$V = \sigma_g^2 \psi + \sigma_e^2 I$$

Overall phenotypic variance-covariance matrix = genetic + error

Obtain z scores and p-values of the effect based on this.

GCTA
a tool for Genome-wide Complex Trait Analysis

GCTA | SMR | GSMR | OSCA | CTG forum | Yang Lab

- Overview
- Download
- FAQ
- Basic Options
- GREML
- GWAS Analysis
  - MLMA
  - fastGWA

fastGWA

fastGWA: A fast MLM-based Genome-Wide Association tool

fastGWA is an ultra-efficient tool for mixed linear model (MLM)-based GWAS analysis of biobank-scale data such as the UK Biobank (see Jiang et al. *Nature Genetics* 2019 for details of the method). Credits: Longda Jiang (method, simulation and analysis), Zhili Zheng (method, software and analysis) and Jian Yang (method and overseeing).

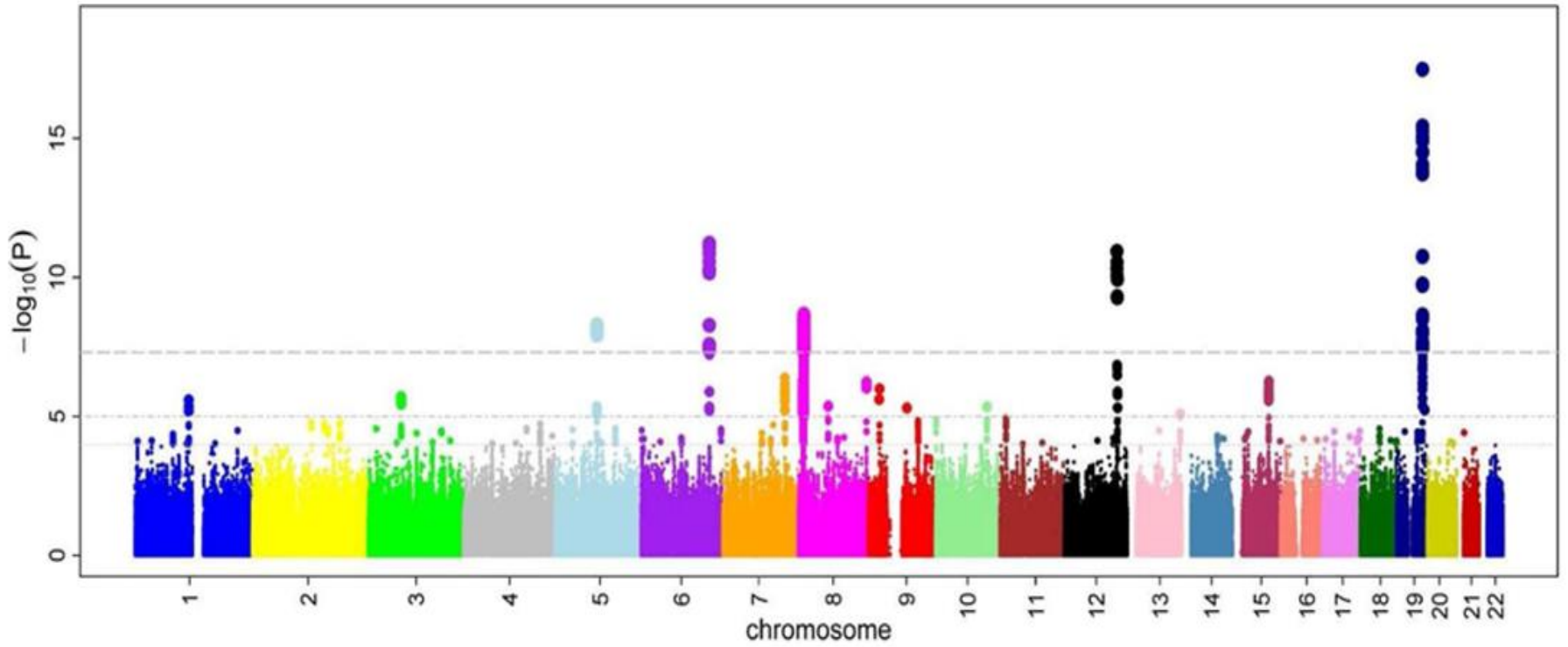
We have applied fastGWA to 2,173 traits on 456,422 array-genotyped and imputed individuals and 2,048 traits on 49,960 whole-exome-sequenced (WES) individuals in the UK Biobank. All the summary statistics are available

Check more recent approaches:
SAIGE (Zhou et al 2018 *Nat Genet*),
REGENIE (Mbatchou et al 2021 *Nat Genet*)

Jiang et al 2019 *Nat Genet*

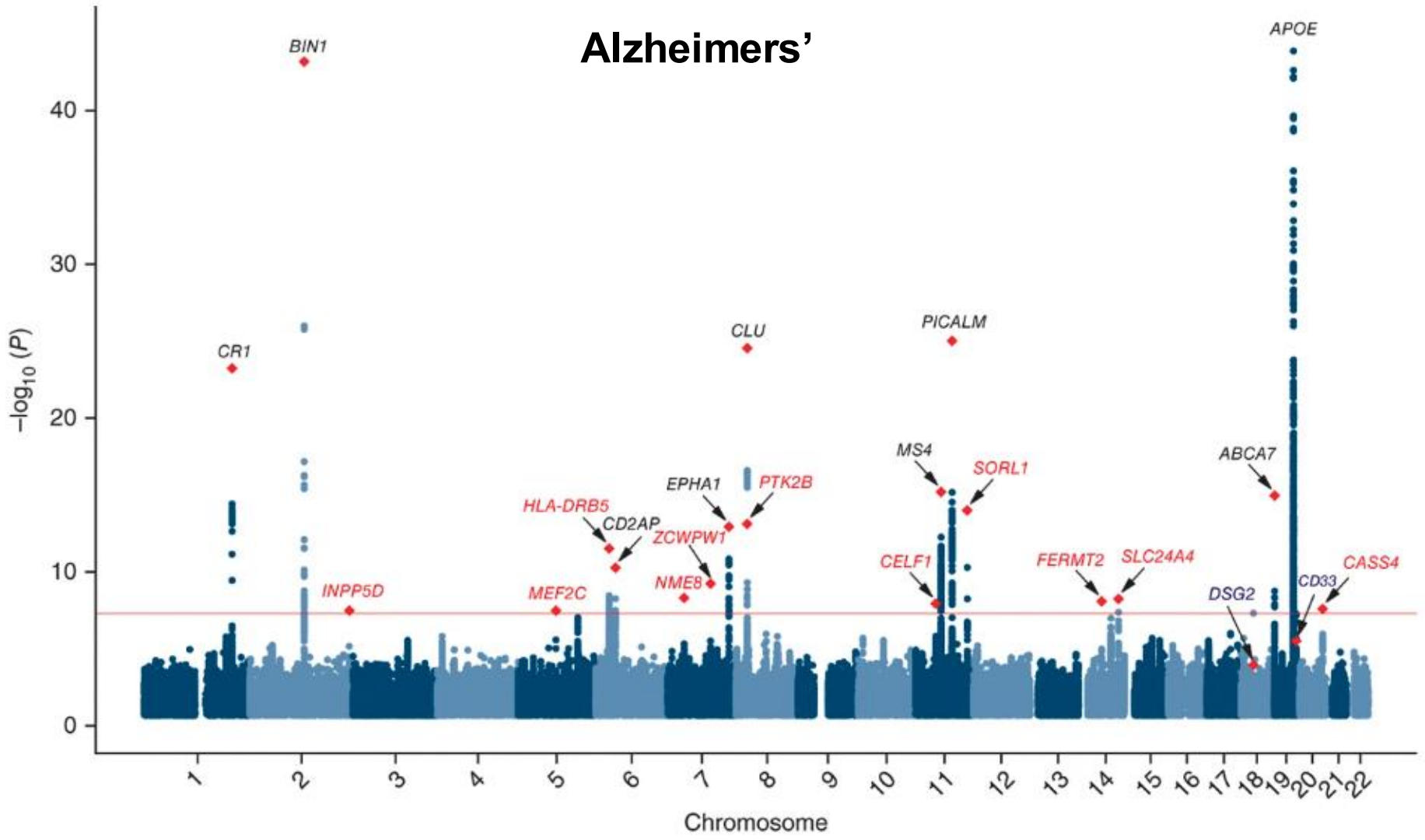
Standard visualization technique for GWAS results

Schizophrenia



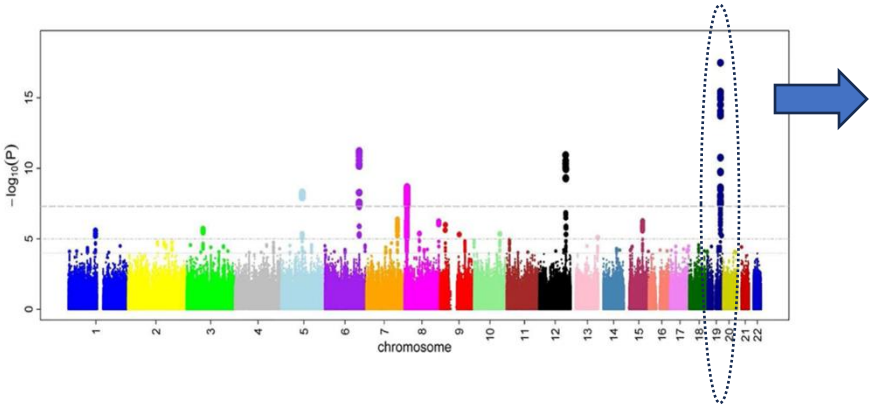
SCZ WG of PGC Consortium, 2014, *Nature*

Standard visualization technique for GWAS results

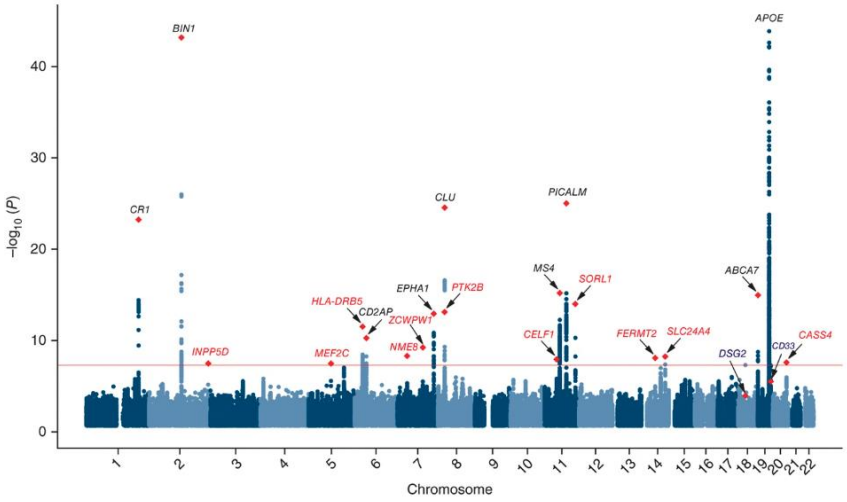


Lambert et al 2013, *Nat Genet*

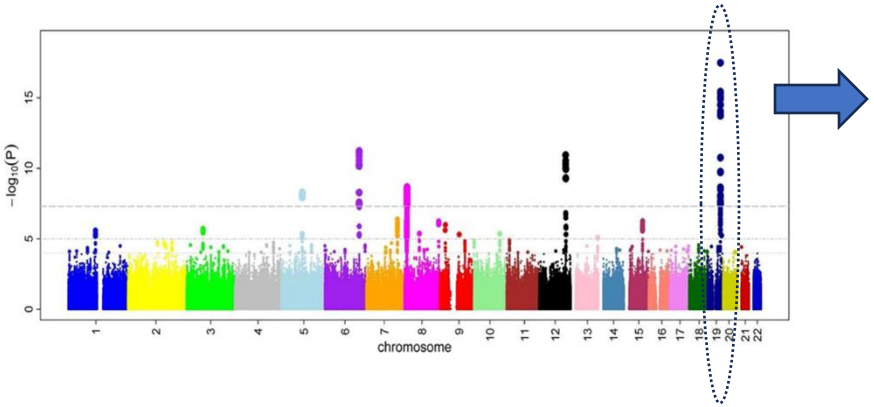
What is common between these GWAS-es?



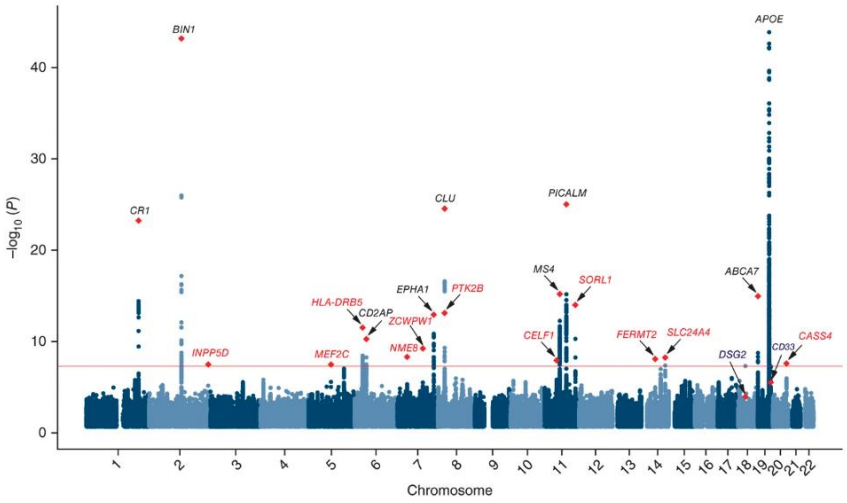
GWAS hits occur in clusters of variants all showing significant effects in same region – this is because of high linkage disequilibrium.



What is common between these GWAS-es?



GWAS hits occur in clusters of variants all showing significant effects in same region – this is because of high linkage disequilibrium.



GWAS signals are highly polygenic encompassing many genes.

We are likely missing out many weaker GWAS effect signals due to stringent p-value thresholds.

Can genotypes explain phenotypic variance across individuals?

Heritability: Proportion of phenotypic variance that can be attributed to genetic effects

Heritability of GWAS hits (h^2_{GWAS}): Squared correlation between best fit linear model of all GWAS hits and the phenotype

$$\max_{w}[\ r^2\ (\sum_{s \in GWAS\ hits} w_s X_{ns}, Y_n)\ ]$$

Heritability of all SNPs (h^2_g): Squared correlation between best fit linear model of all SNPs and the phenotype

$$\max_{w}[\ r^2 (\sum_{s} w_s X_{ns}, Y_n)\ ]$$

Can genotypes explain phenotypic variance across individuals?

Heritability: Proportion of phenotypic variance that can be attributed to genetic effects

Heritability of GWAS hits (h^2_{GWAS}): Squared correlation between best fit linear model of all GWAS hits and the phenotype

$$\max_{w}[\ r^2\ (\sum_{s \in GWAS\ hits} w_s X_{ns}, Y_n)\ ]$$

Heritability of all SNPs (h^2_g): Squared correlation between best fit linear model of all SNPs and the phenotype

$$\max_{w}[\ r^2 (\sum_{s} w_s X_{ns}, Y_n)\ ]$$

How can we estimate SNP heritability h_g^2 from full GWAS association summary statistics

$$\mathbf{Y} = \underbrace{\mathbf{XB}}_{\text{fixed effects}} + \underbrace{u}_{\text{random effects}} + \varepsilon$$

$\mathbf{Y} = N$ x 1 vector of phenotypes

$\mathbf{X} = N$ x c matrix of c covariates

$\mathbf{B} = c$ x 1 vector of effect sizes of c covariates

$u \sim \mathrm{N}(0,\sigma_g^2\mathbf{A})$ is residual variance due to genetic effects

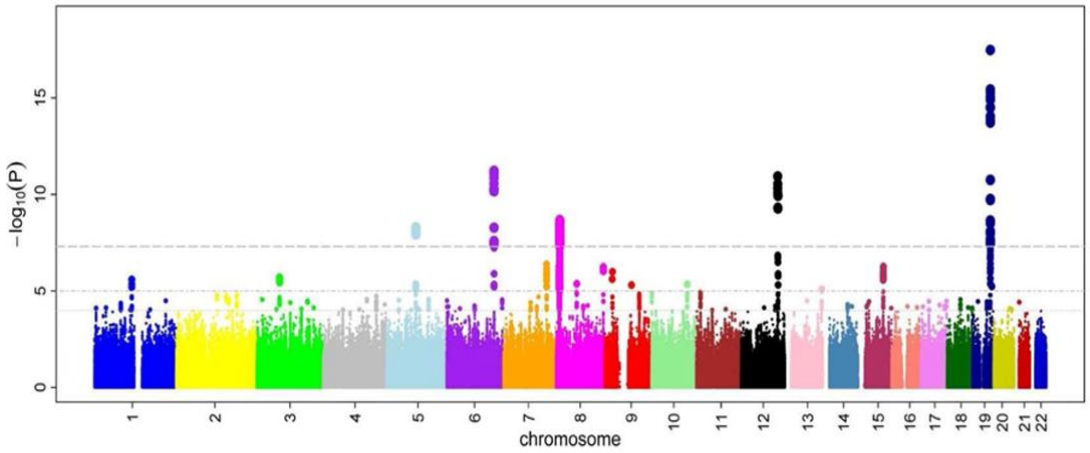
$\varepsilon \sim \mathrm{N}(0,\sigma_e^2\mathbf{I})$ is residual variance due to environmental effects

$\mathbf{V} = \mathrm{Var}(u + \varepsilon) = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$

$$\boxed{\text{heritability } h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)}$$

Yang et al 2010 *Nat Genet*
Kang et al 2010 *Nat Genet*
Zaitlen et al 2013 *PLoS Genet*

There is a *big gap* between only focusing on GWAS hits and looking at all of the GWAS association summary



Schizophrenia

$0.07 < 0.24$

h^2_{GWAS} h^2_g

Hidden Heritability

Lichtenstein et al 2009 *Lancet*
Lee et al. 2012 *Nat Genet*
Trubetskoy et al. 2022 *Nature*

This gap has been largely resolved for Adult height GWAS

A saturated map of common genetic variants associated with human height

Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira, Heather H. Highland, Yingjie Ji, Tugce Karaderi, Kuang Lin, Kreete Lüll, Deborah E. Malden, 23andMe Research Team, VA Million Veteran Program, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), Lifelines Cohort Study, The PRACTICAL Consortium, Understanding Society Scientific Group, ... Joel N. Hirschhorn

+ Show authors

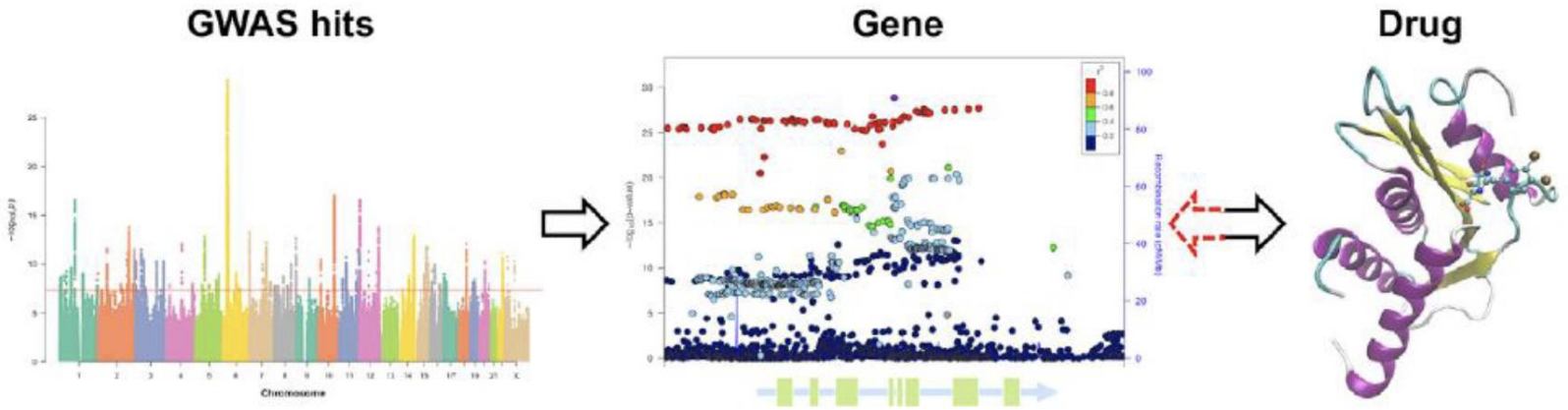
Nature (2022) | Cite this article

"Here, using data from a genome-wide association study of 5.4 million individuals of diverse ancestries, we show that 12,111 independent SNPs that are significantly associated with height account for nearly all of the common SNP-based heritability."

Also see O'Connor et al 2021 *Nat Genet*

Clinical and therapeutic implications of GWAS

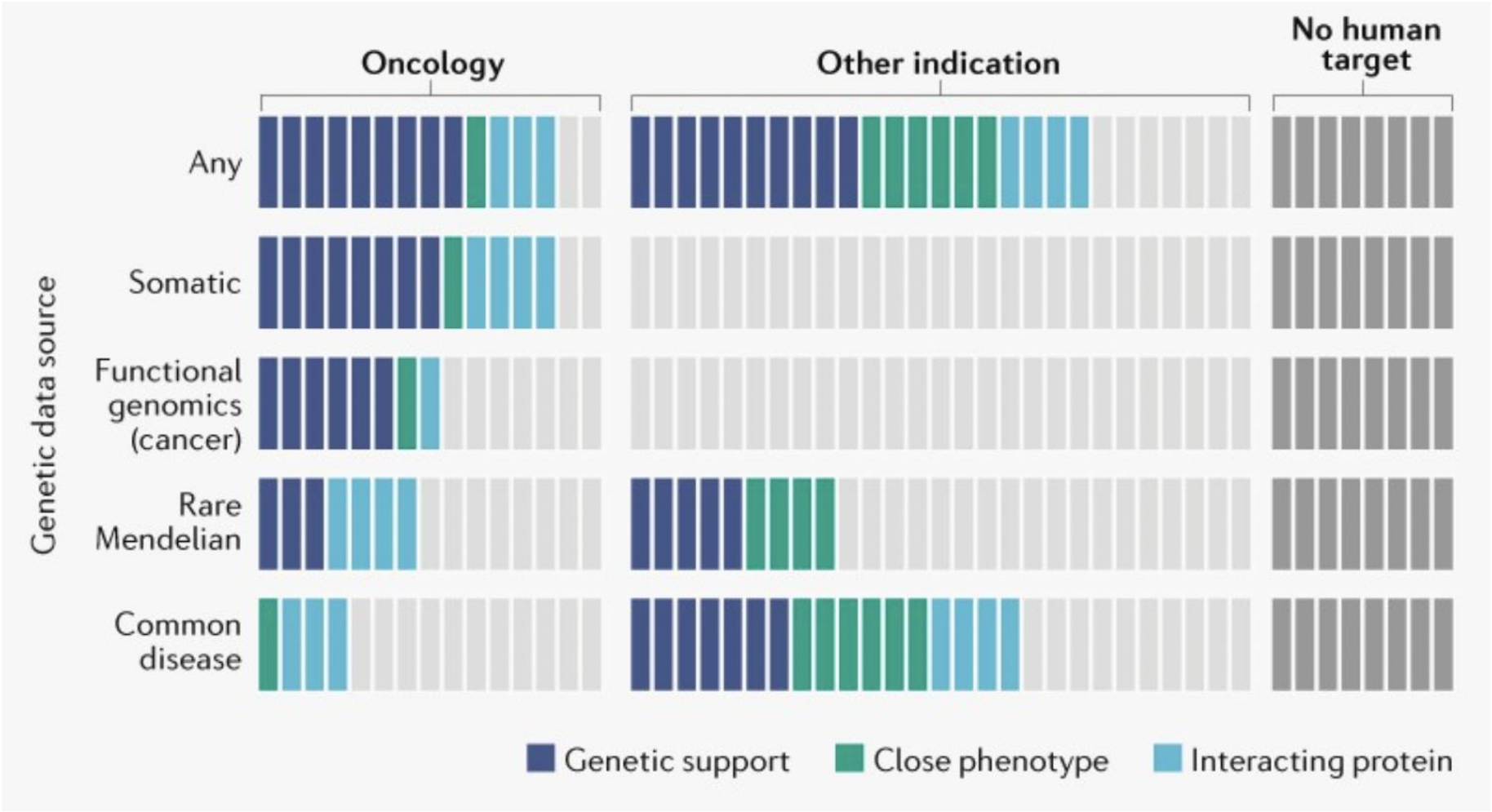
Is GWAS actually important? (GWAS hits to drugs)



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

Visscher et al 2017 *AJHG*
Fang et al 2019 *Nat Genet*

Is GWAS actually important? (GWAS hits to drugs)

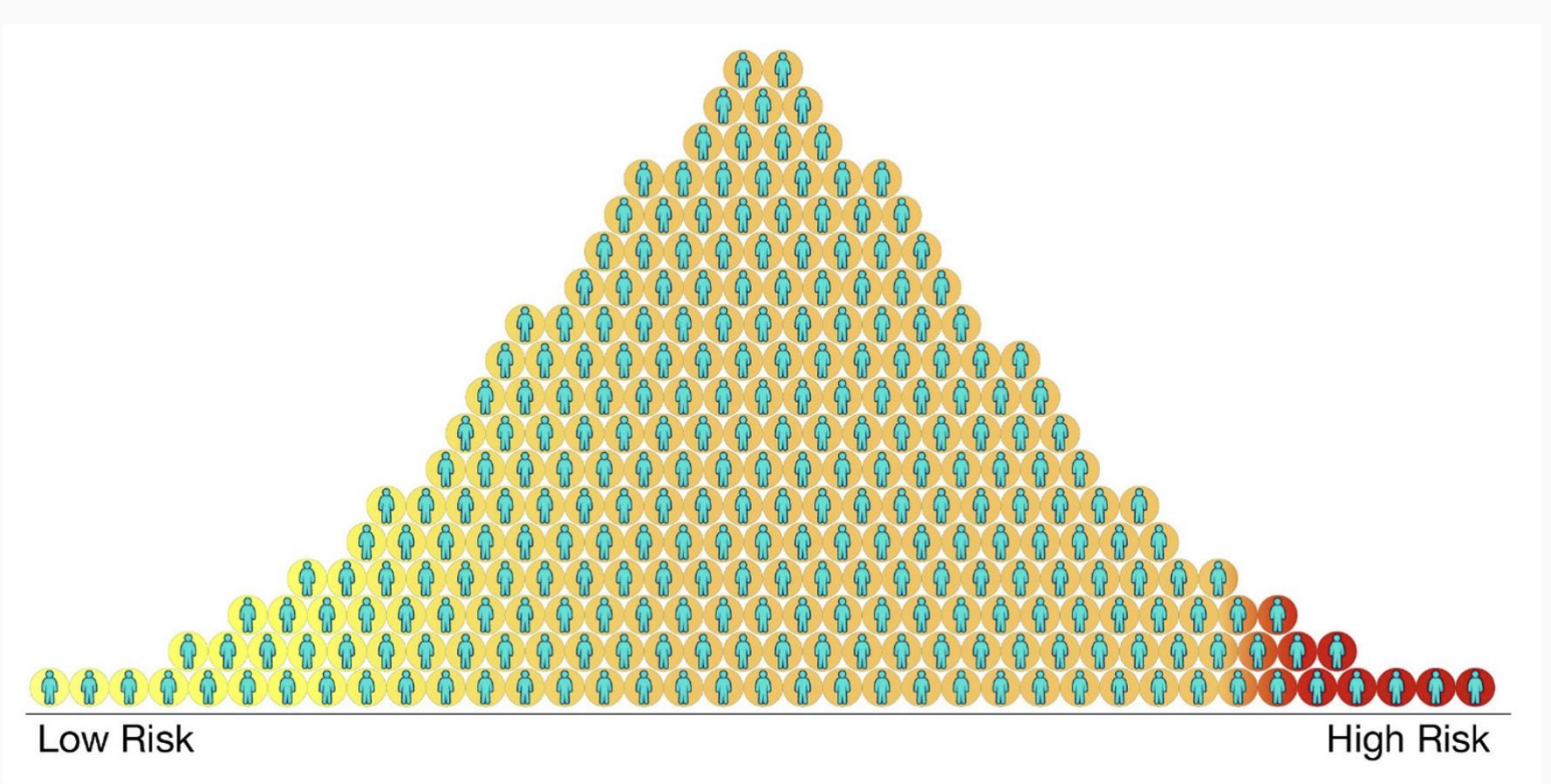


33 of 50 FDA approved drugs in 2021 have genetic support, with highest implicated from common disease GWAS.

Ochoa et al 2022 *Nat Rev Drug Disc.*

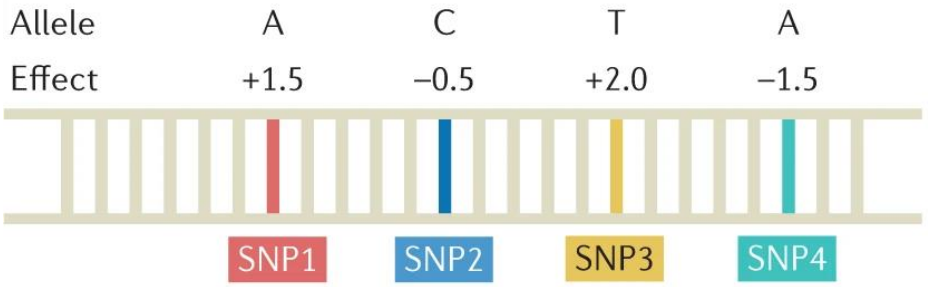
Is GWAS actually important? (Genetic risk score)

Identify the genetic risk for any individual for diseases and traits based on their genetic make-up (genotypes across all SNPs). Are they at risk for a specific disease?



How to calculate polygenic risk scores?

① GWAS summary statistics



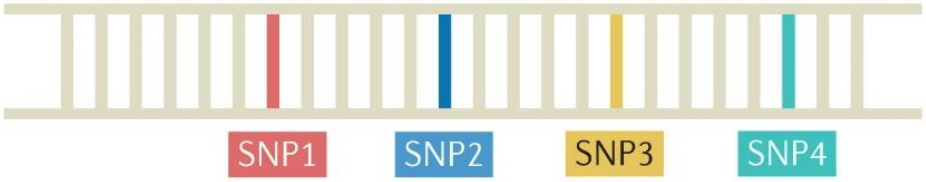
② Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

How to calculate polygenic risk scores?

1 GWAS summary statistics

Allele	A	C	T	A
Effect	+1.5	–0.5	+2.0	–1.5



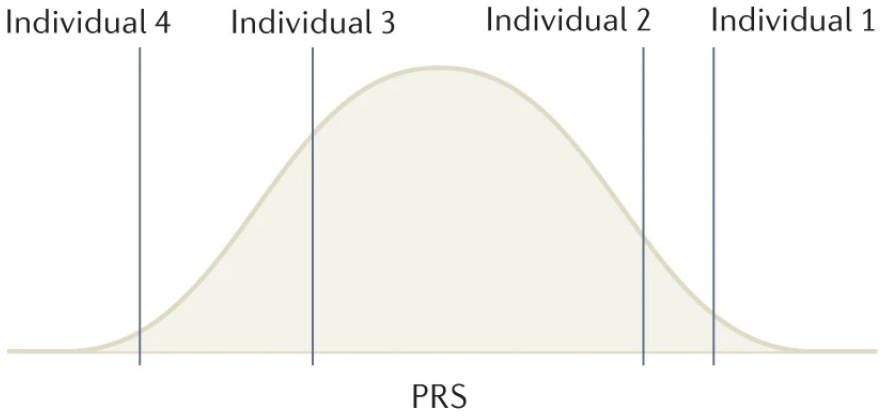
2 Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

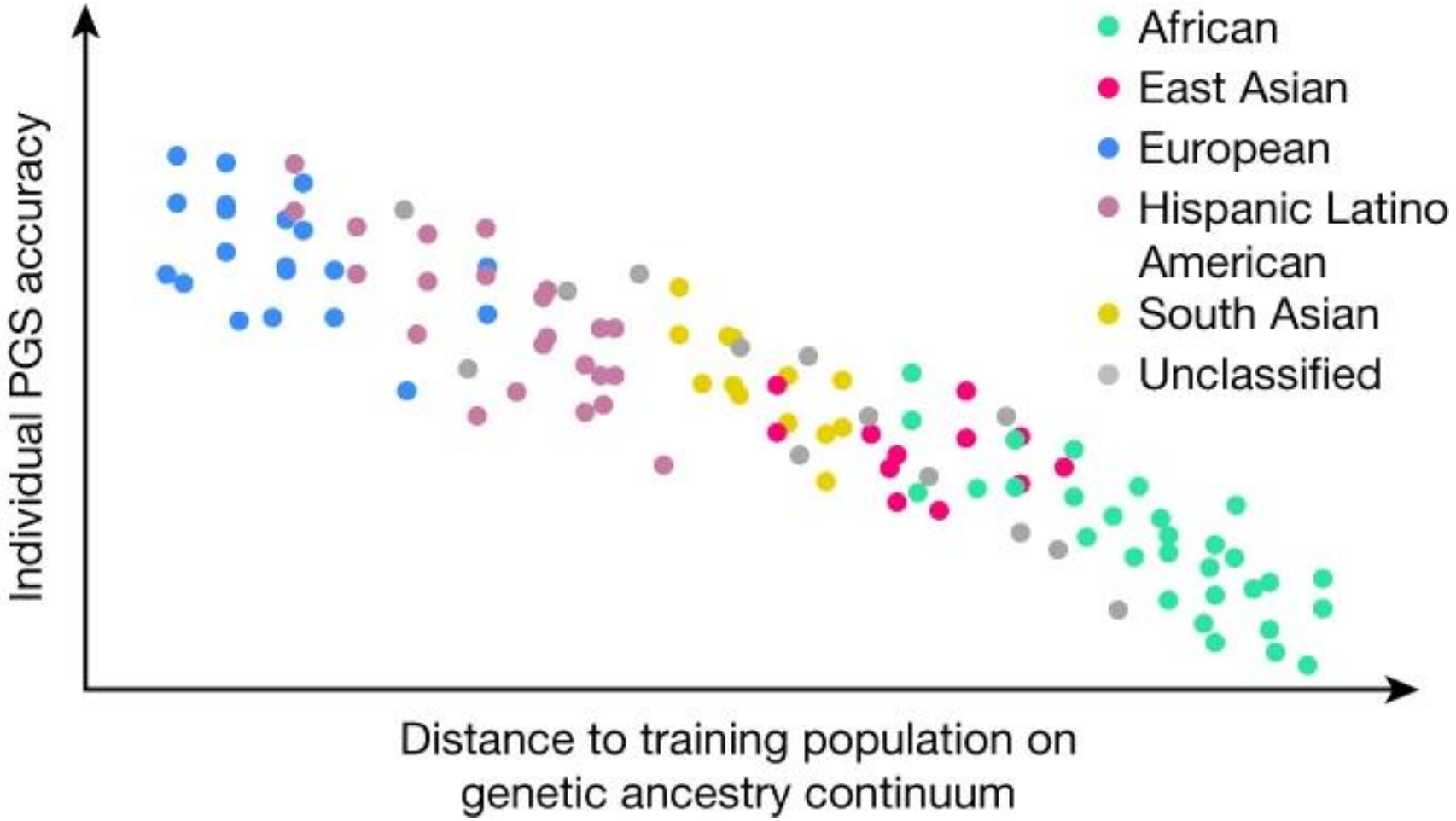
3 Polygenic risk score

Individual 1	1.5	–	0.5	+	4.0	–	0.0	=	**5.0**
Individual 2	1.5	–	0.0	+	2.0	–	1.5	=	**2.0**
Individual 3	0.0	–	1.0	+	2.0	–	1.5	=	**–0.5**
Individual 4	0.0	–	1.0	+	0.0	–	3.0	=	**–4.0**

4 PRS distribution

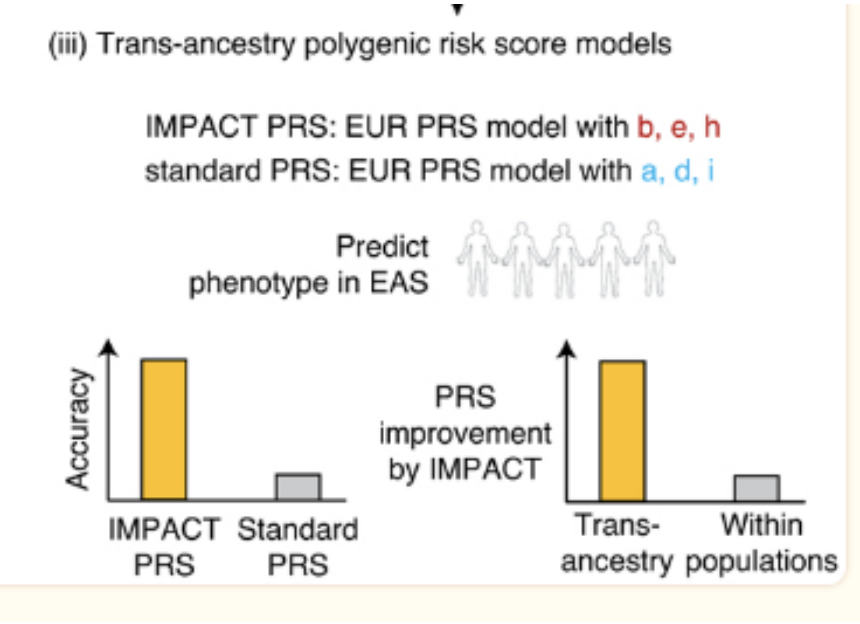


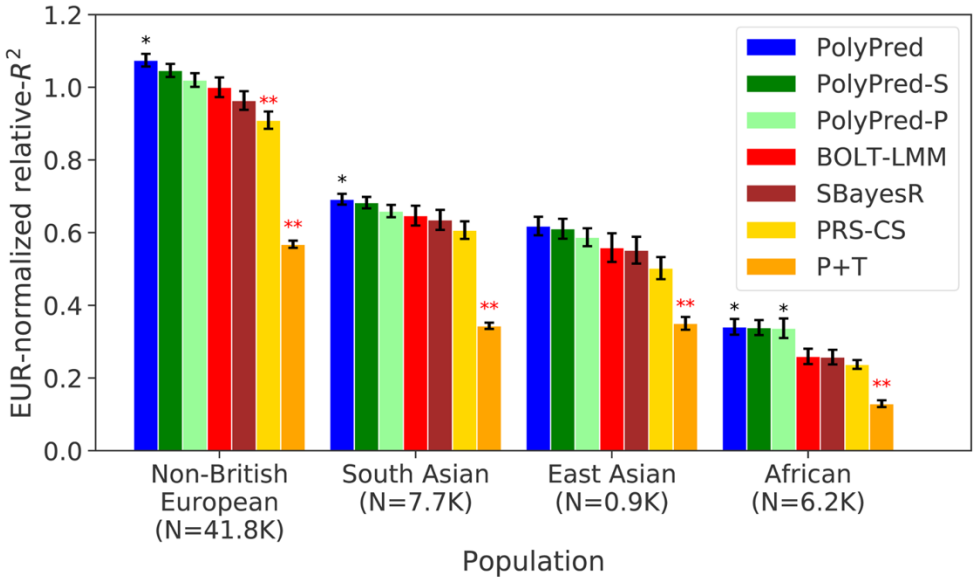
A big challenge in polygenic risk scores (representation)



Ding et al 2023 *Nature*

Integrating functional annotations can resolve polygenic risk score transferability across populations





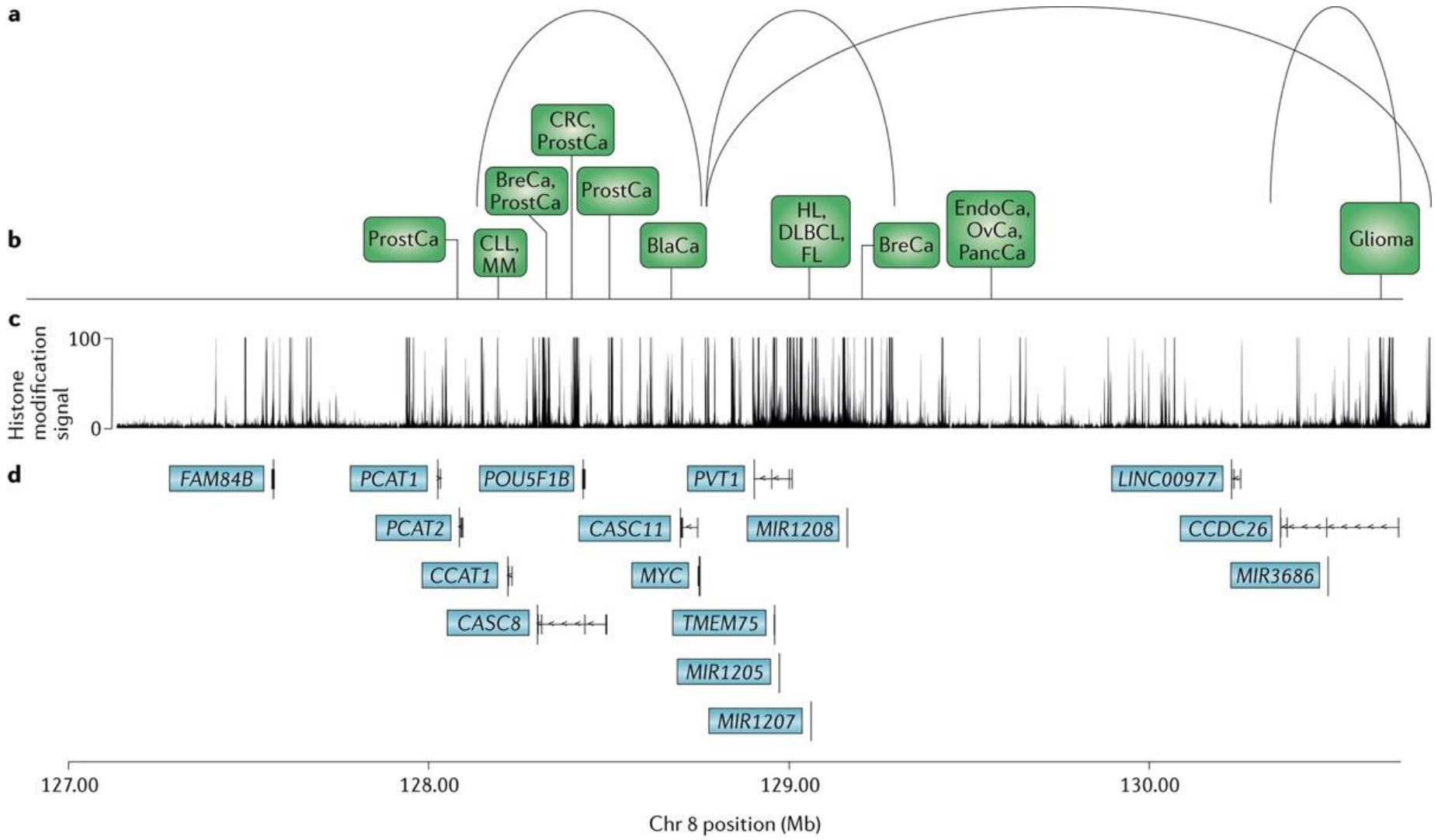
IMPACT: integrative combination of TF binding functional predictions

Amariuta..Dey et al 2020 *Nat Genet*

PolyPred: integrative combination of functional annotations to drive trans-ancestry prediction risk accuracy

Weissbrod et al 2022 *Nat Genet*

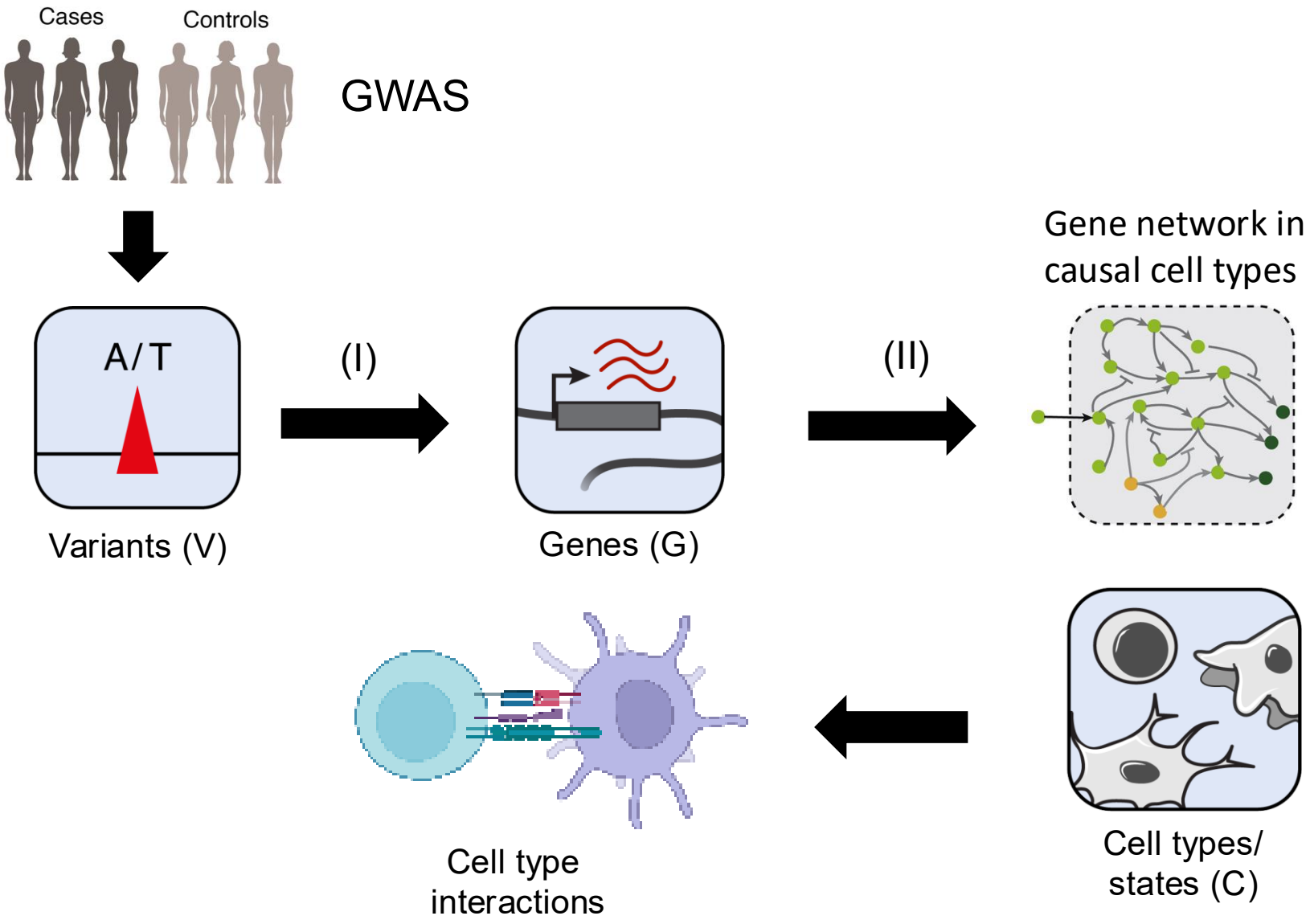
Pleiotropic associations in Genome-wide associations



Cancer-related pleiotropy

Learn shared associations across traits:
MTAG (Turley et al 2018 *Nat Genet*)

Understanding the functional basis of GWAS variants



Claussnitzer et al 2021, *Trends Genet*; Mathieson et al 2021, *AJHG*

GWAS-to-function (Overview)

Linkage disequilibrium can hinder identification of causal variant for both GWAS and eQTL studies

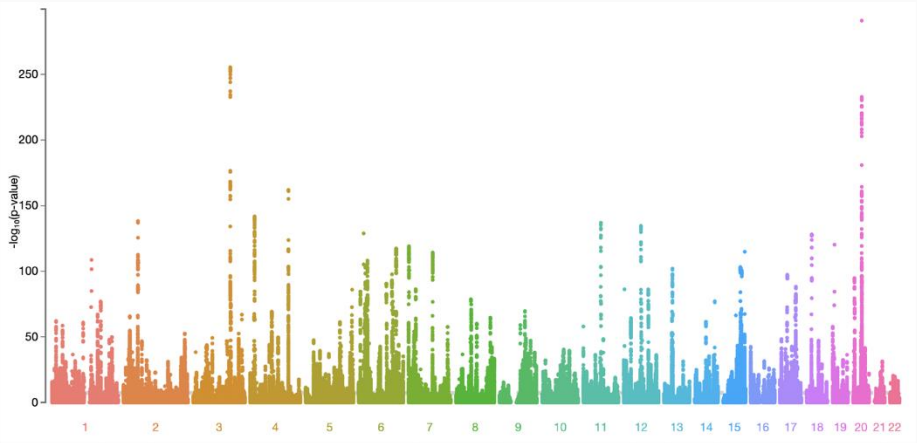
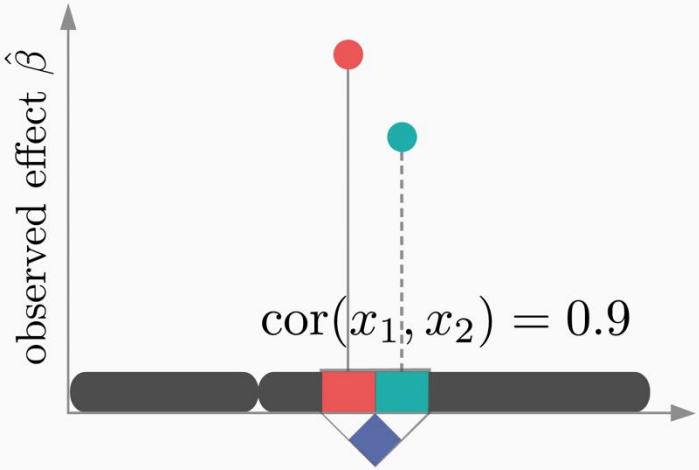


Figure: UK Biobank height GWAS, `http://nealelab.is/uk-biobank`



Linkage disequilibrium can hinder identification of causal variant for both GWAS and eQTL studies

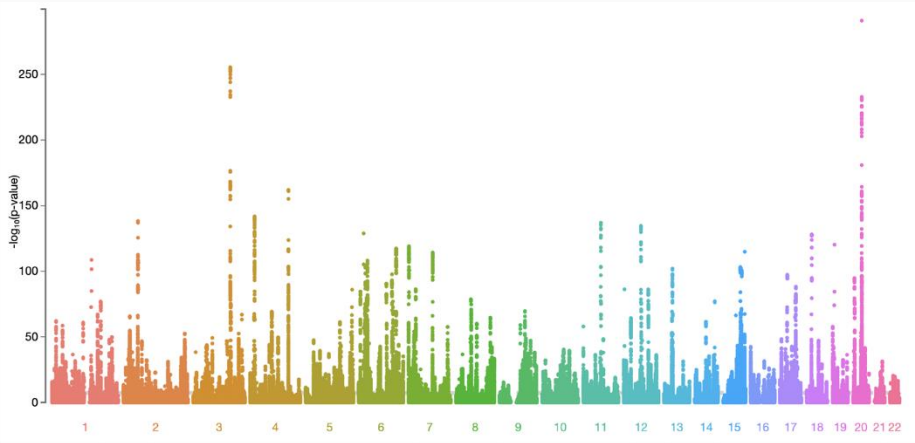
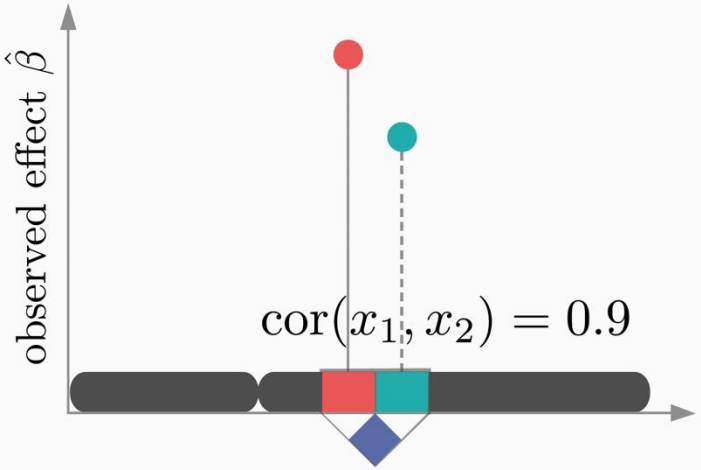
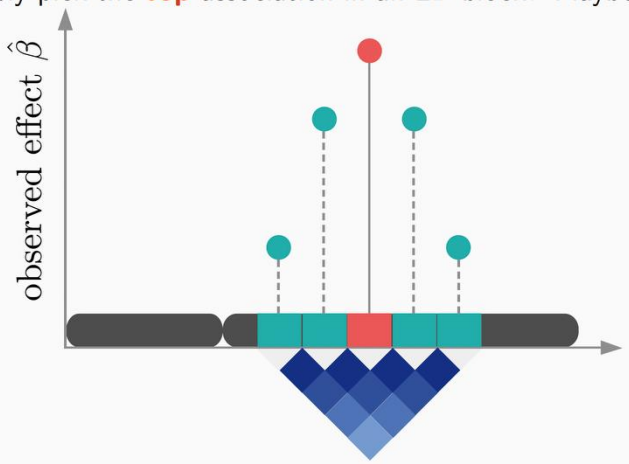


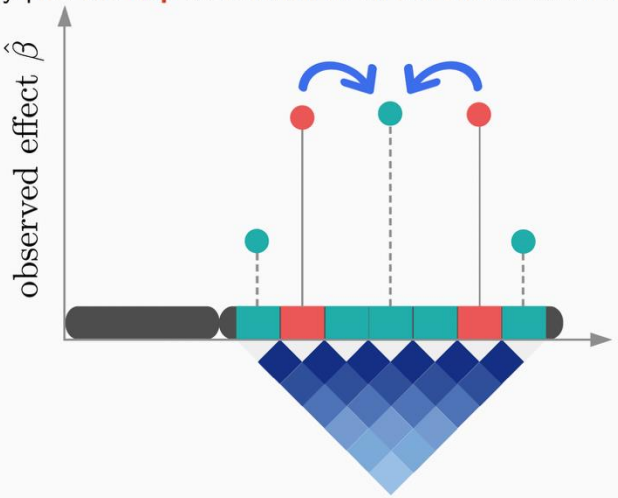
Figure: UK Biobank height GWAS, `http://nealelab.is/uk-biobank`



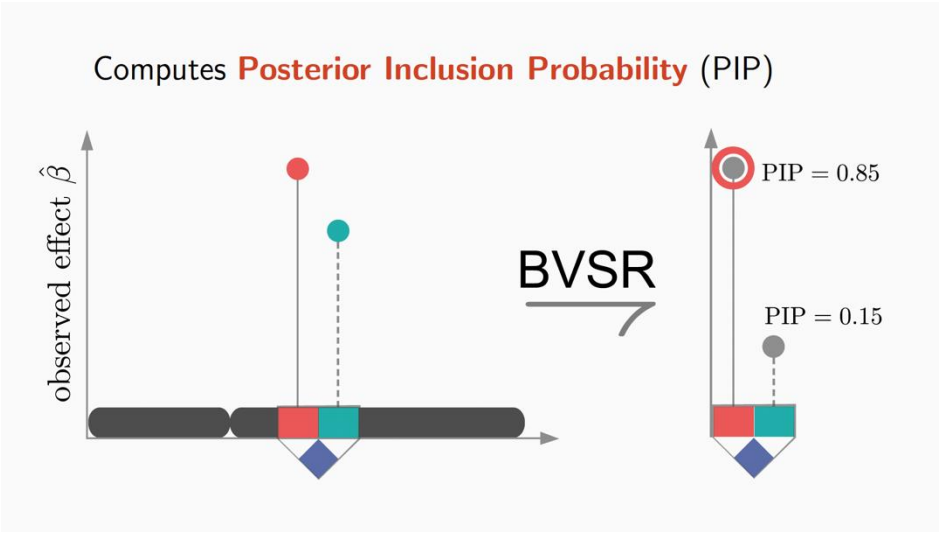
Simply pick the **top** association in an LD block? Maybe?

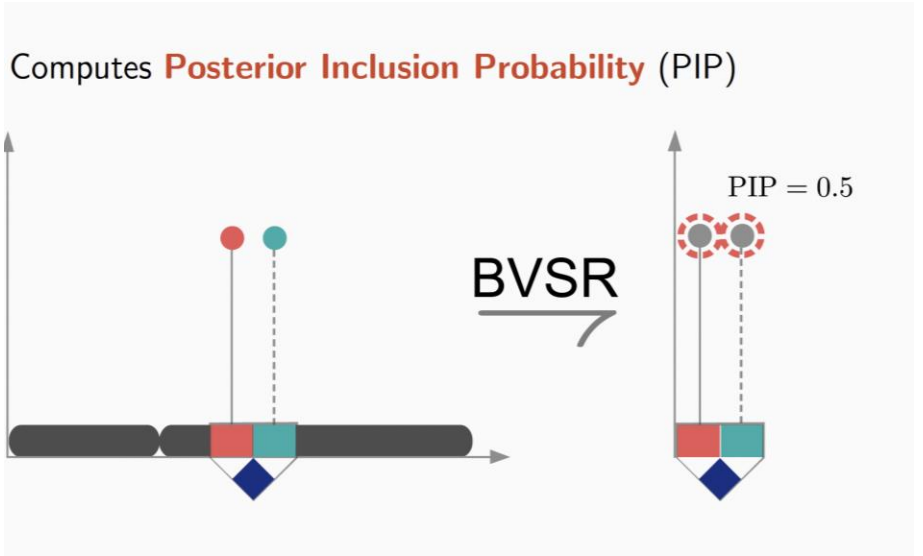


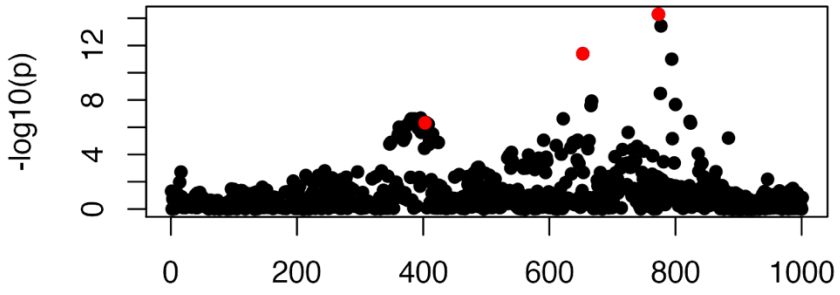
Simply pick the **top** association in an LD block? ... or not!

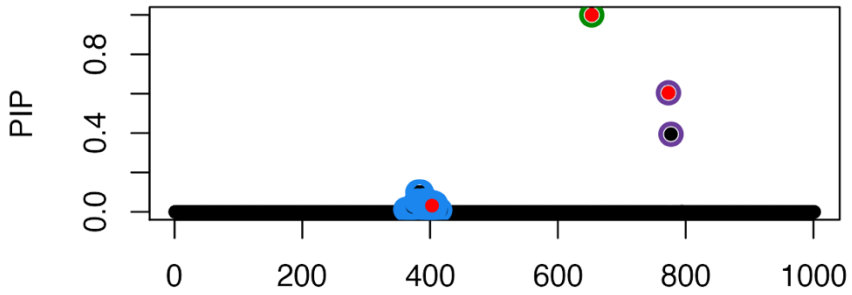


SuSIE: Method to perform Bayesian variable selection to identify independent causal GWAS variants or sets of variants when it is not sure





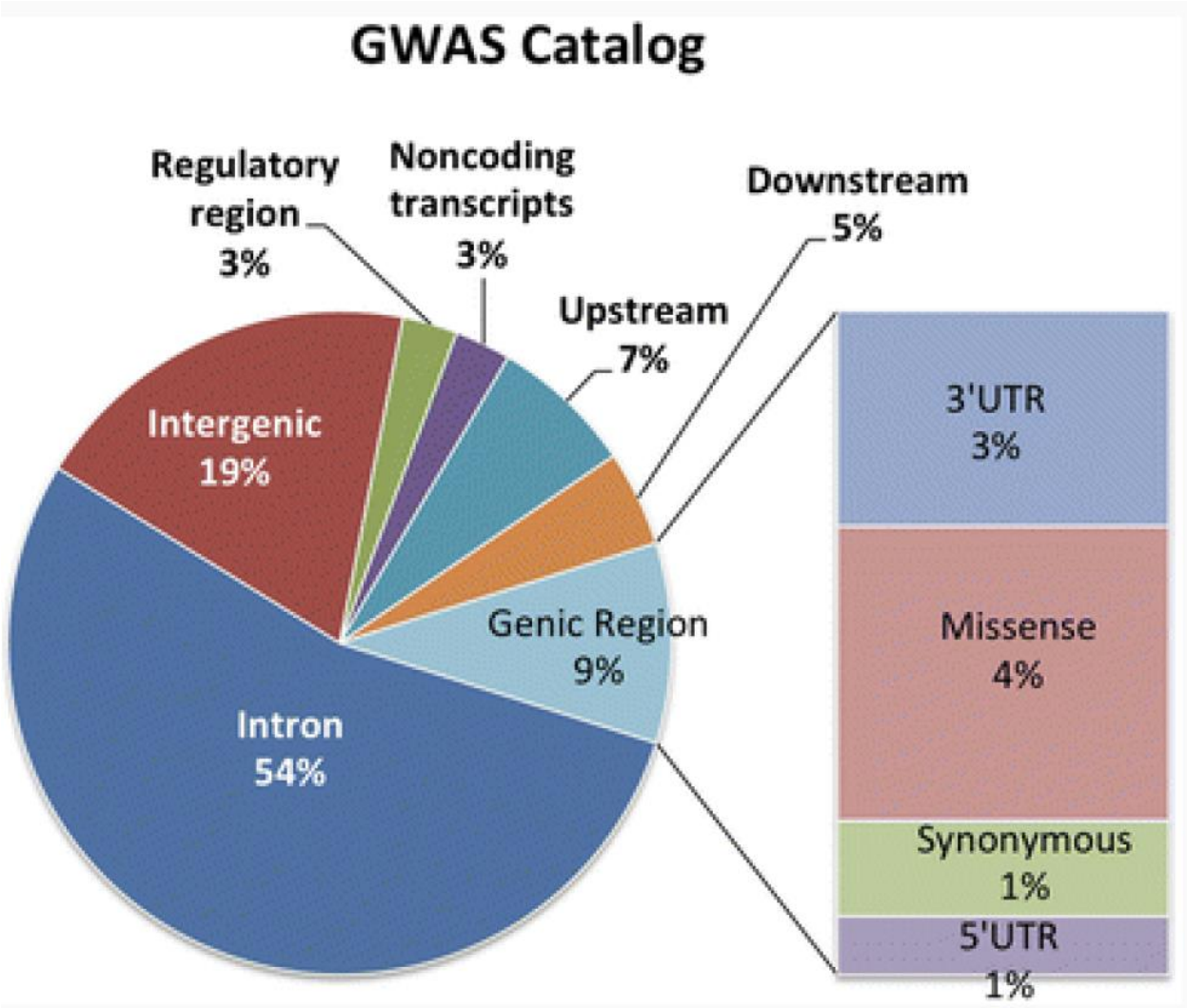




3 colors correspond to 95% **credible sets**: A credible set says a causal variant is within this set with 95% probability

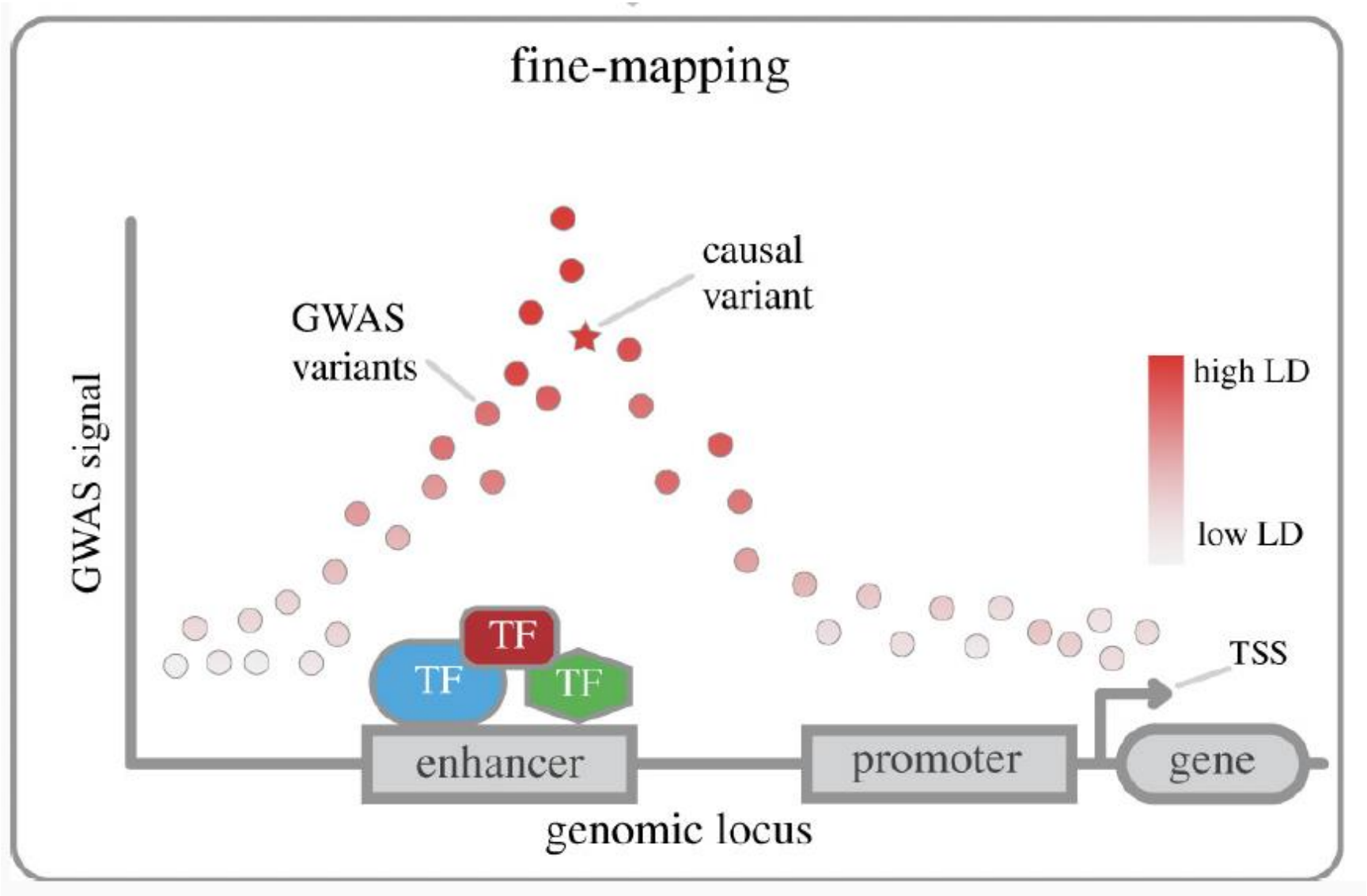
SuSIE: Wang et al 2020 *JRSS-B*

Making sense of the function of GWAS variants



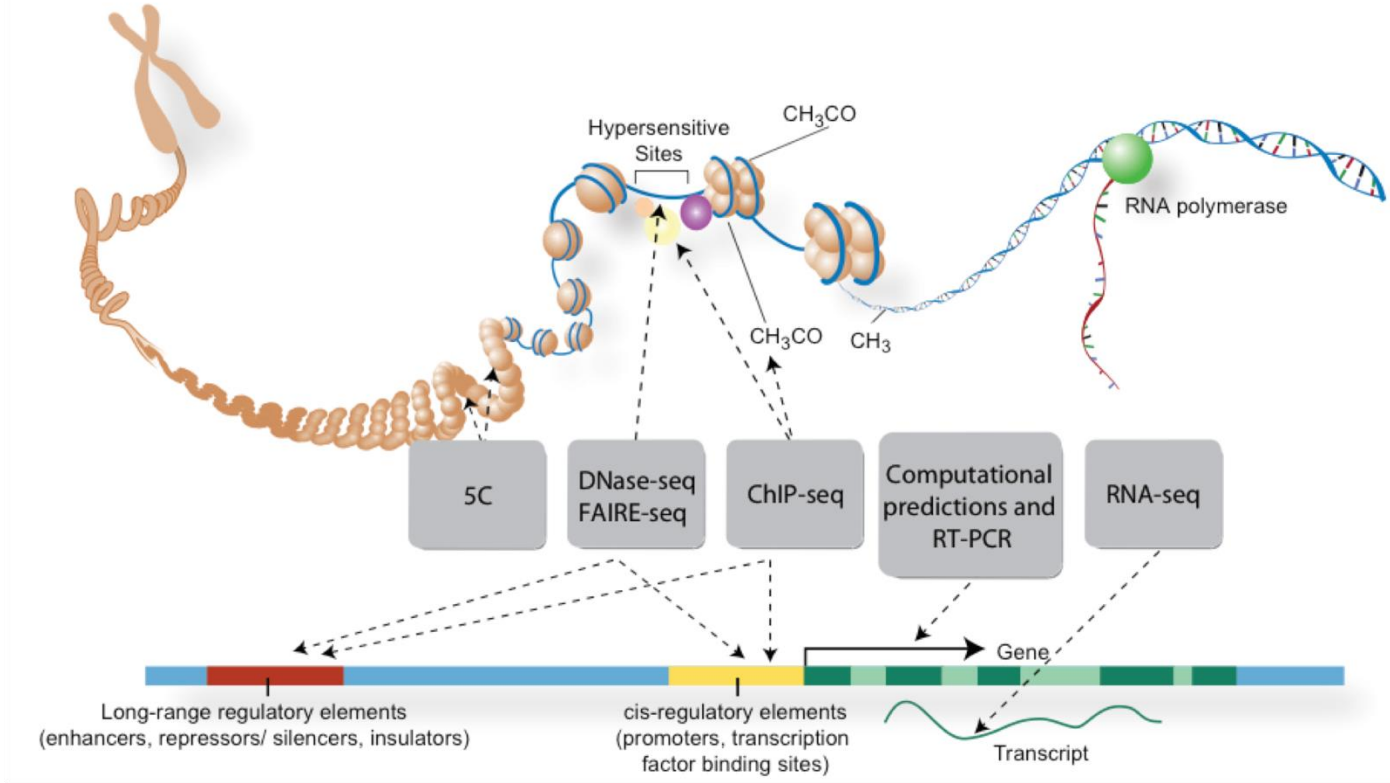
Lee et al 2018 *Human Genetics*

GWAS signals can be confounded by LD. Can we use underlying function to find the causal variant?

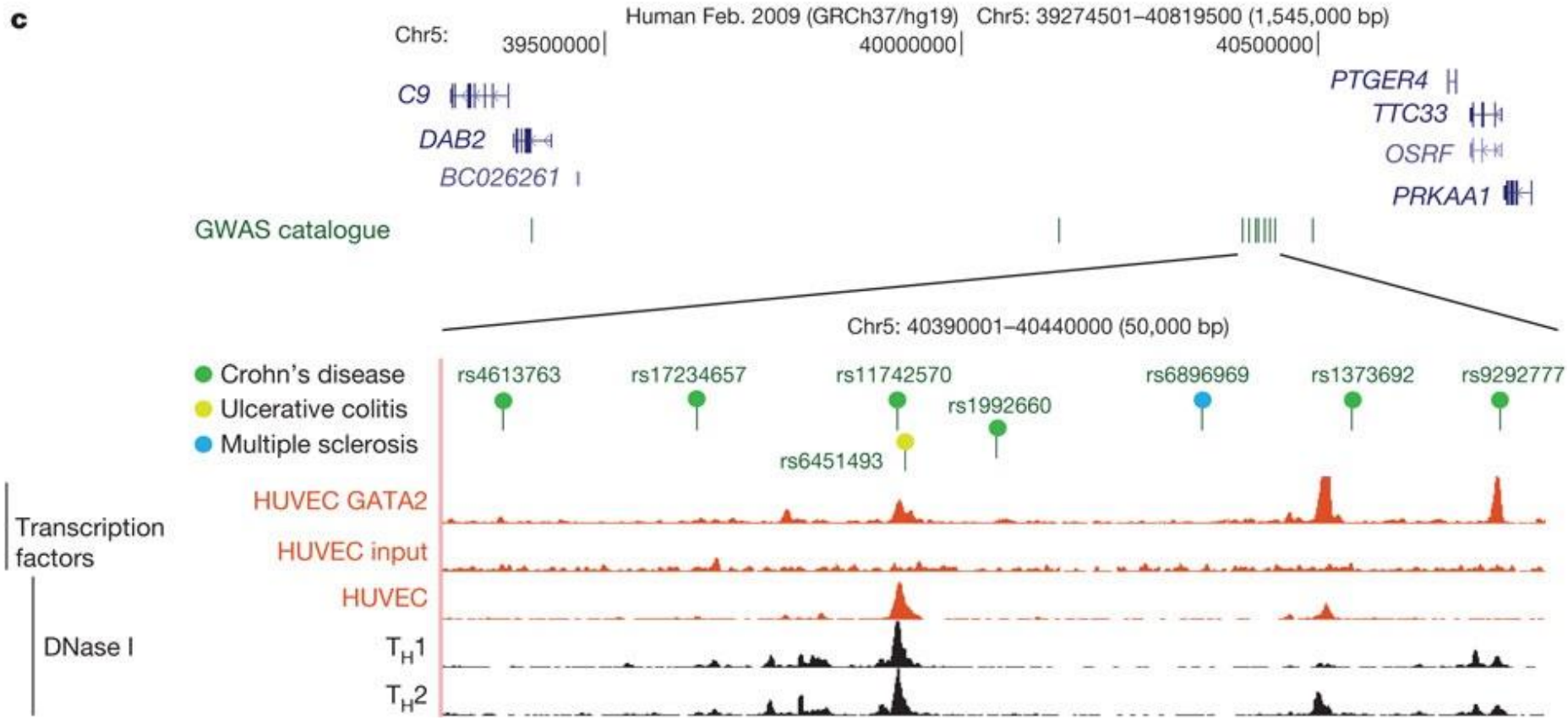


Mapping the function of genomic elements in DNA (ENCODE: 2003-2023)

In 2003, the ENCODE consortium was launched by NIH to study the non-coding regions of the genome and identify functional elements



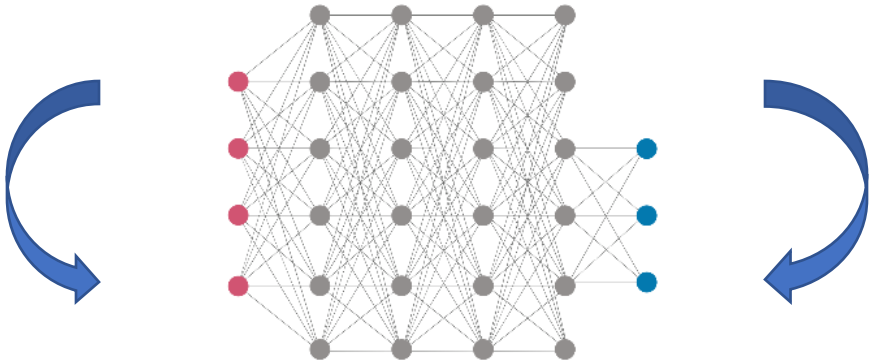
Overlapping genome-wide functional annotation tracks against GWAS disease-associated variants

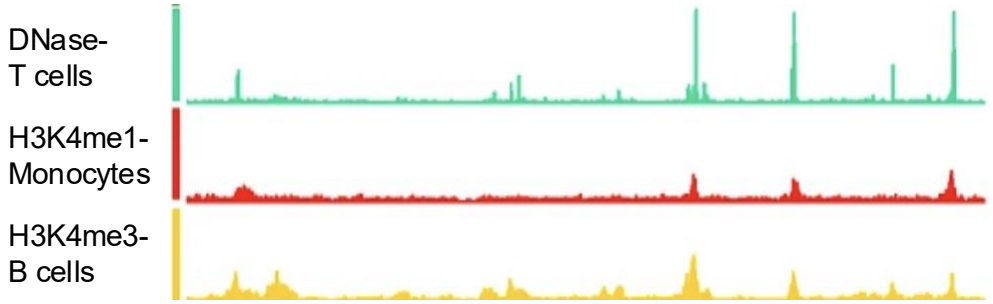


ENCODE consortium 2012, *Nature*

Sequence-based deep learning models trained on epigenomic features

DNA sequences
(1-hot encoding DNA)

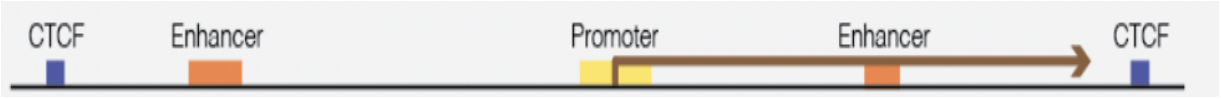


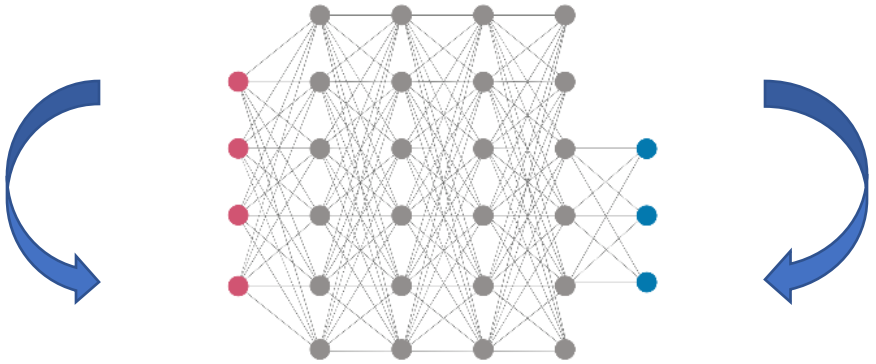


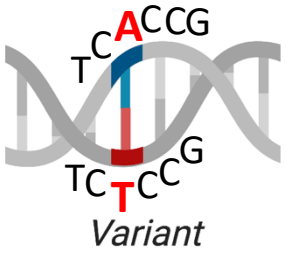
For each sequence, generates a prediction of affinity for each feature f at the site of the sequence.

Sequence-based deep learning models trained on epigenomic features

DNA sequences
(1-hot encoding DNA)



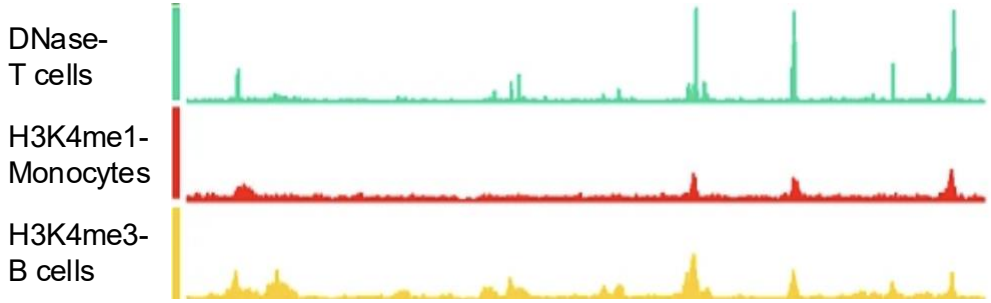




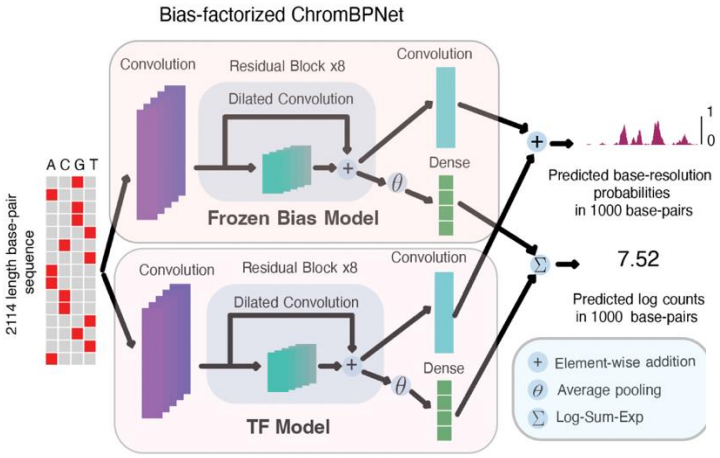
…TCACCG… …TCTCCG…

p^f q^f

$$\Delta^f = | p^f - q^f |$$



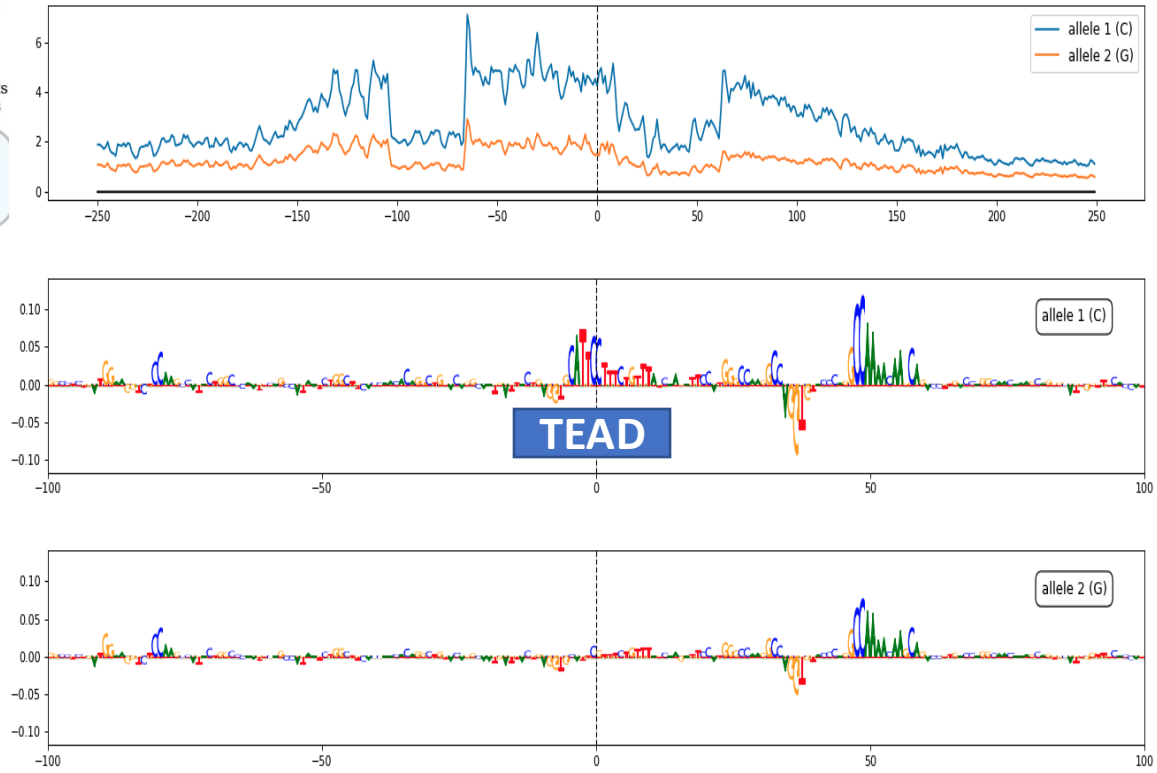
ChromBPNet deep learning model captures sequence mediated function at GWAS variants



https://github.com/kundajelab/chrombpnet

Coronary Artery Disease GWAS variant

rs4266144 (C/G): Human Quiescent SMCs

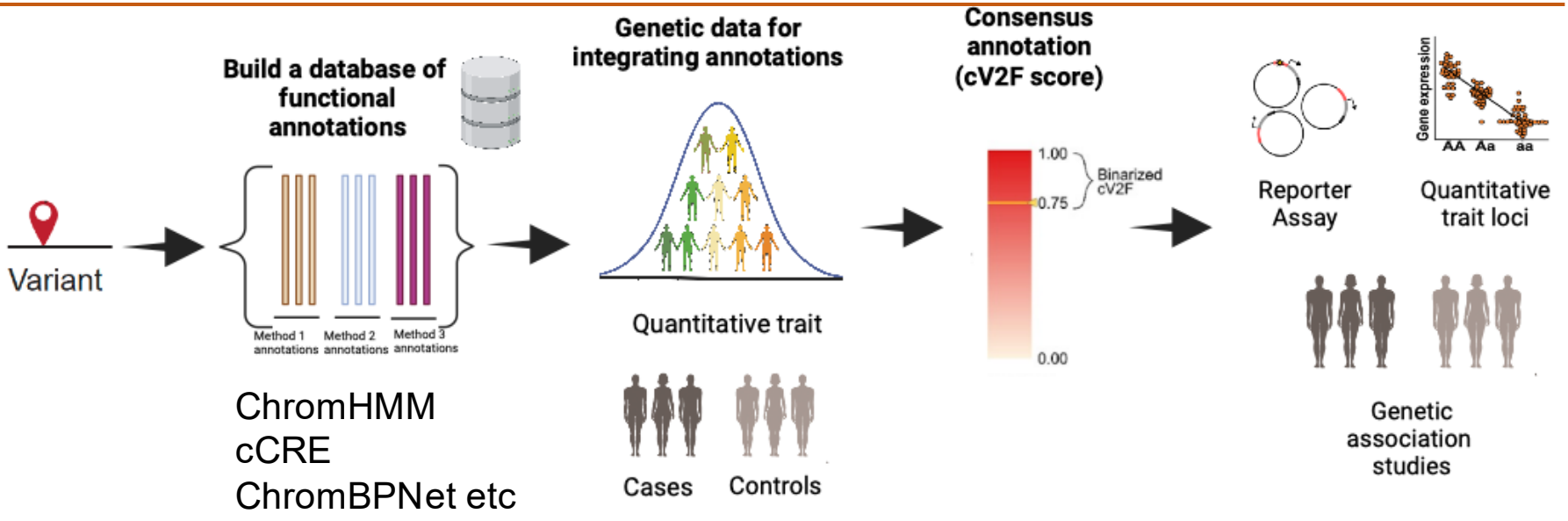


Also see

Enformer: Avsec et al 2021 *Nat Methods*
BPNet: : Avsec et al 2021 *Nat Genet*

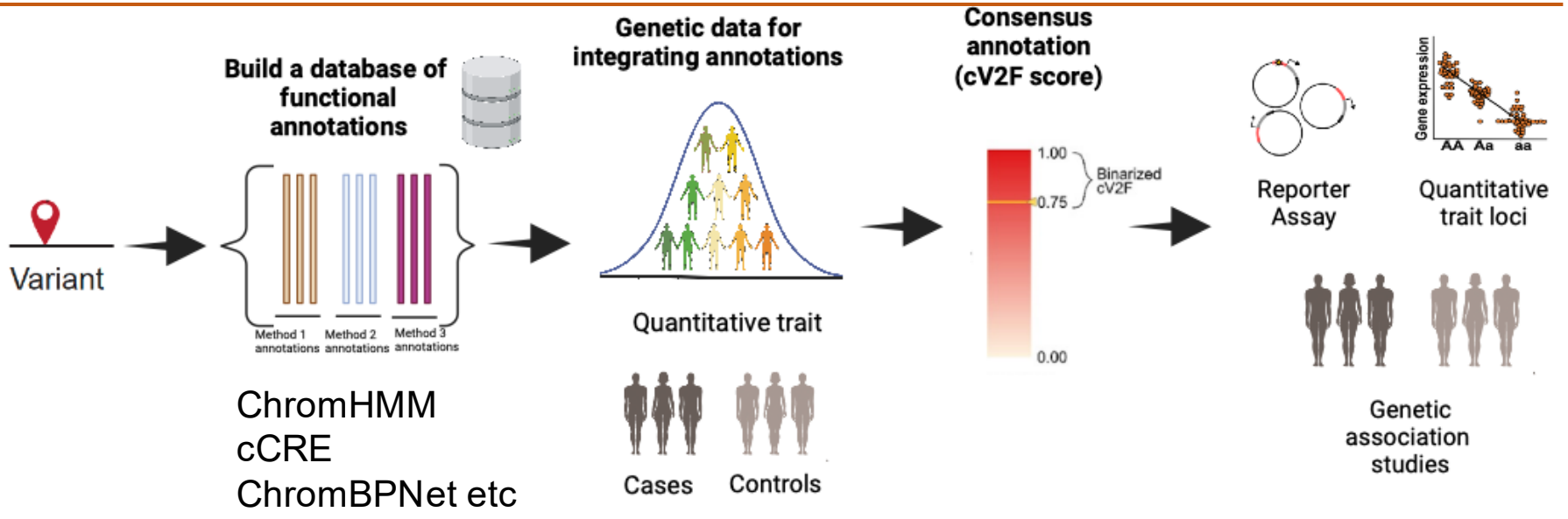
Pampari et al 2024 bioRxiv
Courtesy: Anshul Kundaje, Stanford

Generating an optimal variant-to-disease function score



Fabiha et al 2024 bioRxiv, in rev *Nat Genet*

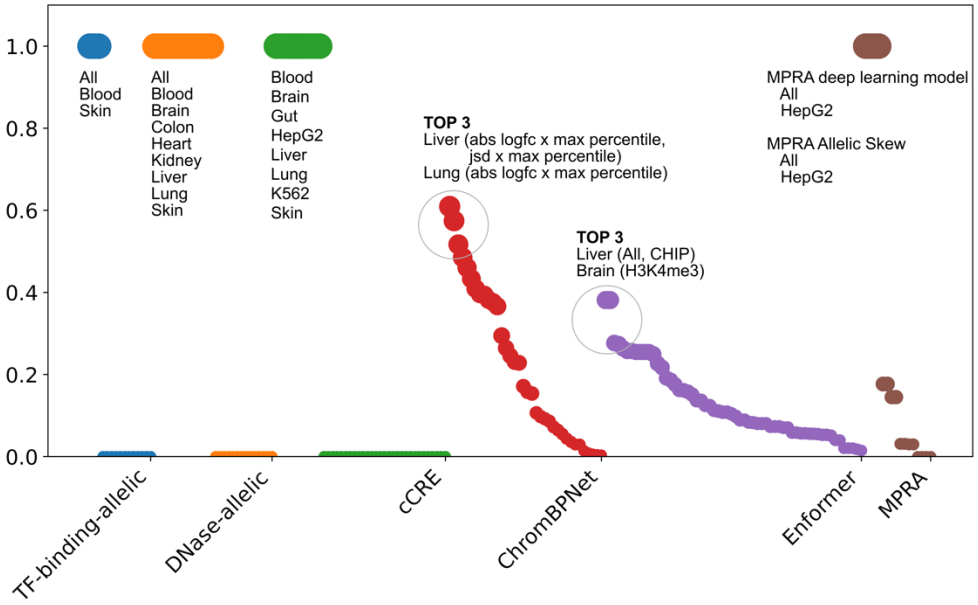
Generating an optimal variant-to-disease function score



rs12740374

cV2F-score = 0.96
cV2F-Liver = 0.93

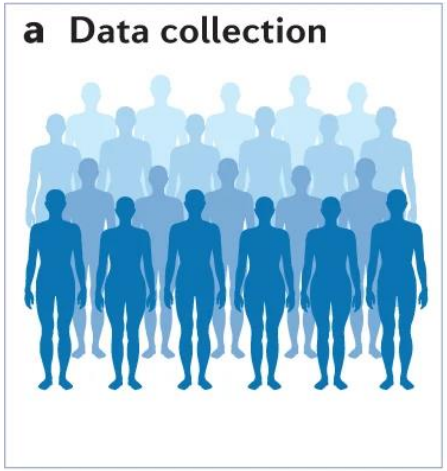
creates a C/EBP transcription factor binding site and alters the hepatic expression of the *SORT1* gene (Musunuru et al 2010 *Nature*)

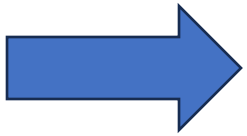


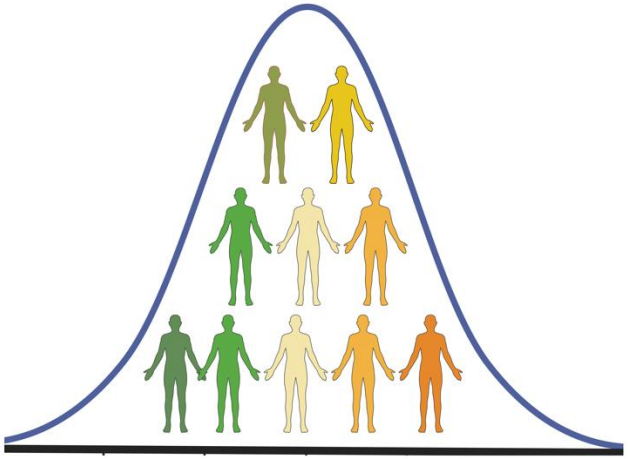
Fabiha et al 2024 bioRxiv, in rev *Nat Genet*

Linking quantitative trait loci to disease GWAS

Tracking genetic variation of gene expression phenotype (eQTL : expression quantitative trait loci)



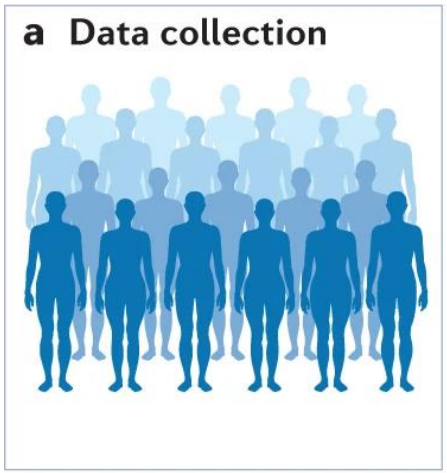


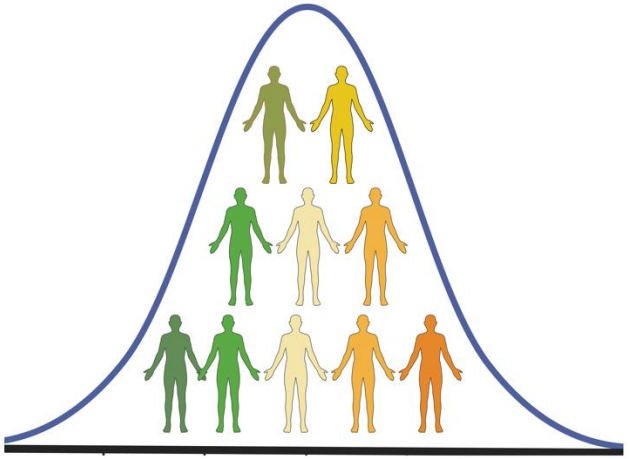


Gene expression for gene G
in tissue/cell-type T

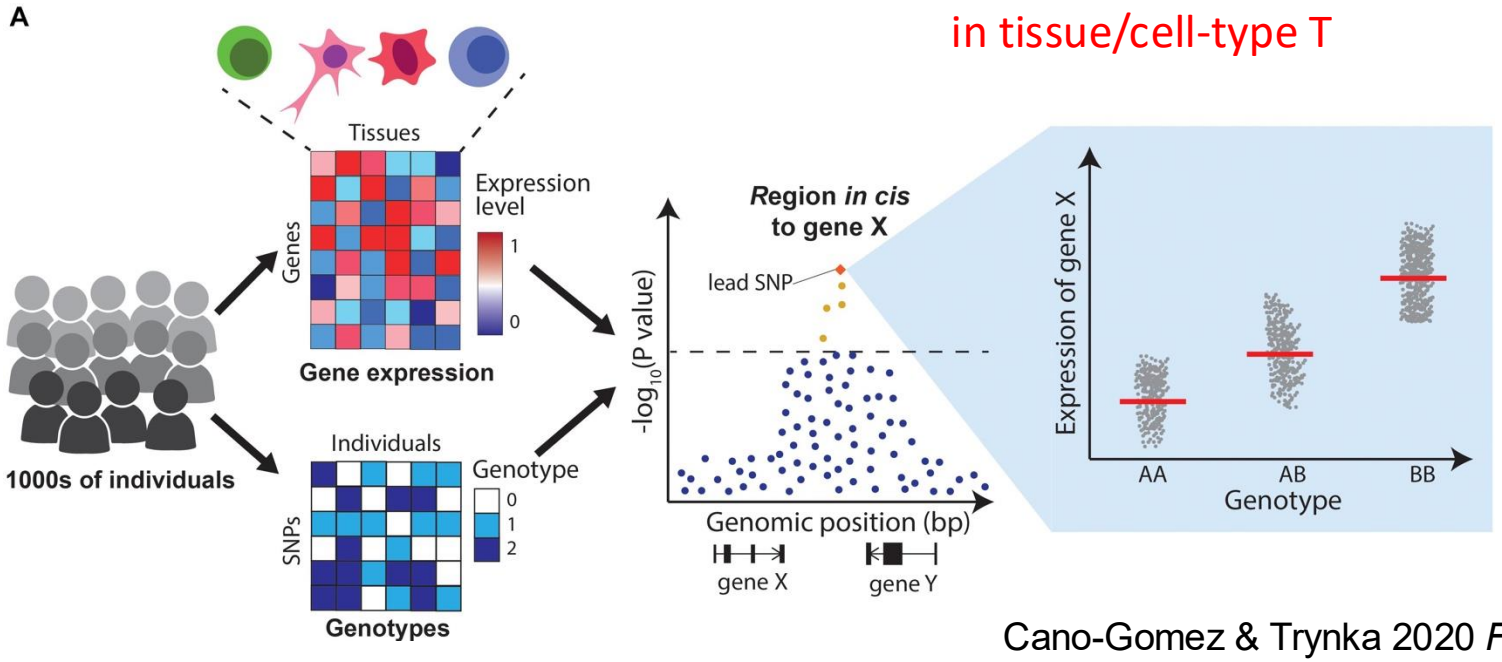
Cano-Gomez & Trynka 2020 *Front Genet*

Tracking genetic variation of gene expression phenotype (eQTL : expression quantitative trait loci)





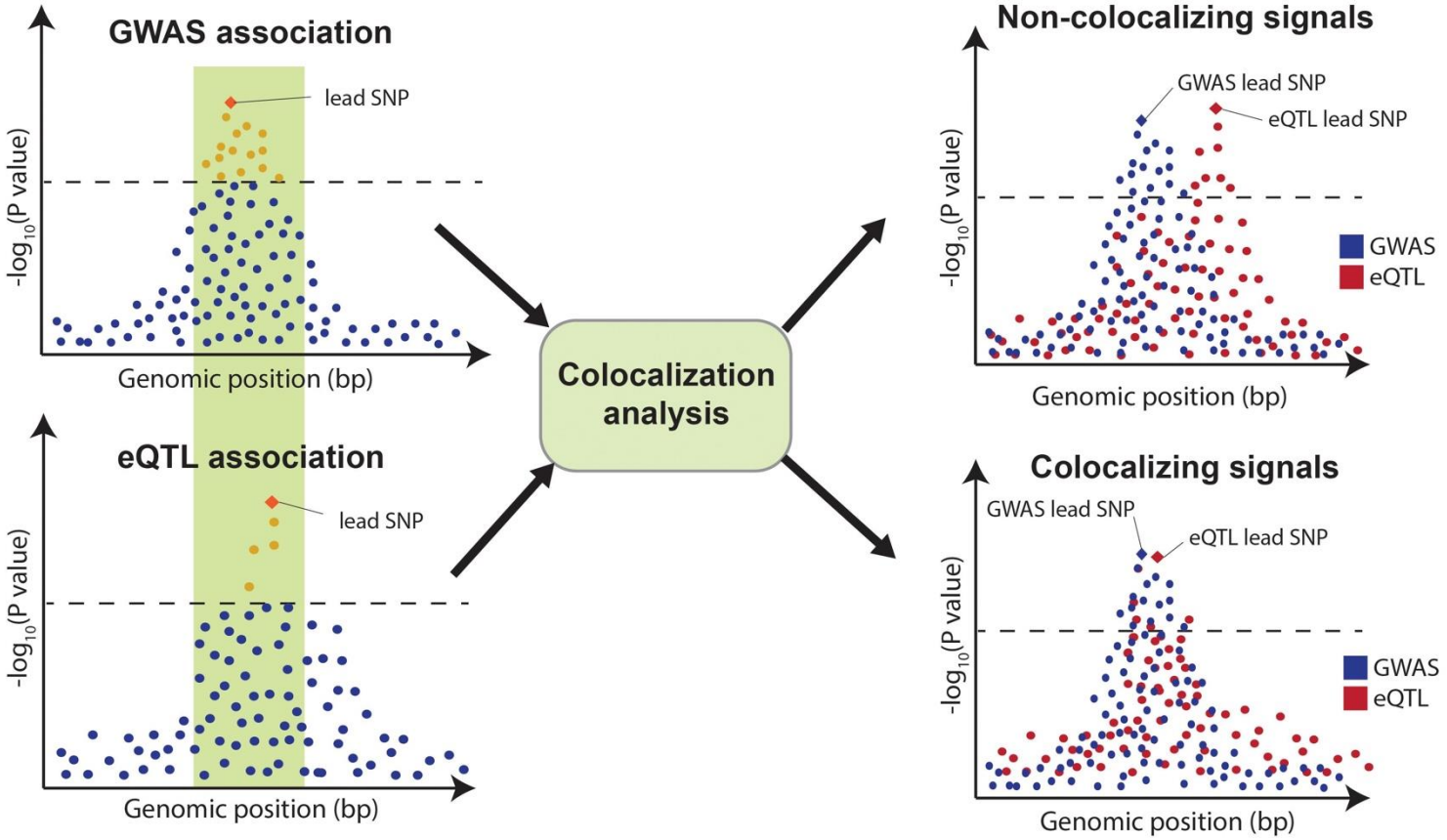
Gene expression for gene G in tissue/cell-type T



Cano-Gomez & Trynka 2020 *Front Genet*

Statistical colocalization: Identifying shared causal variants between a disease trait and an eQTL

Typically performed for one gene and for one tissue separately against one focal disease GWAS.

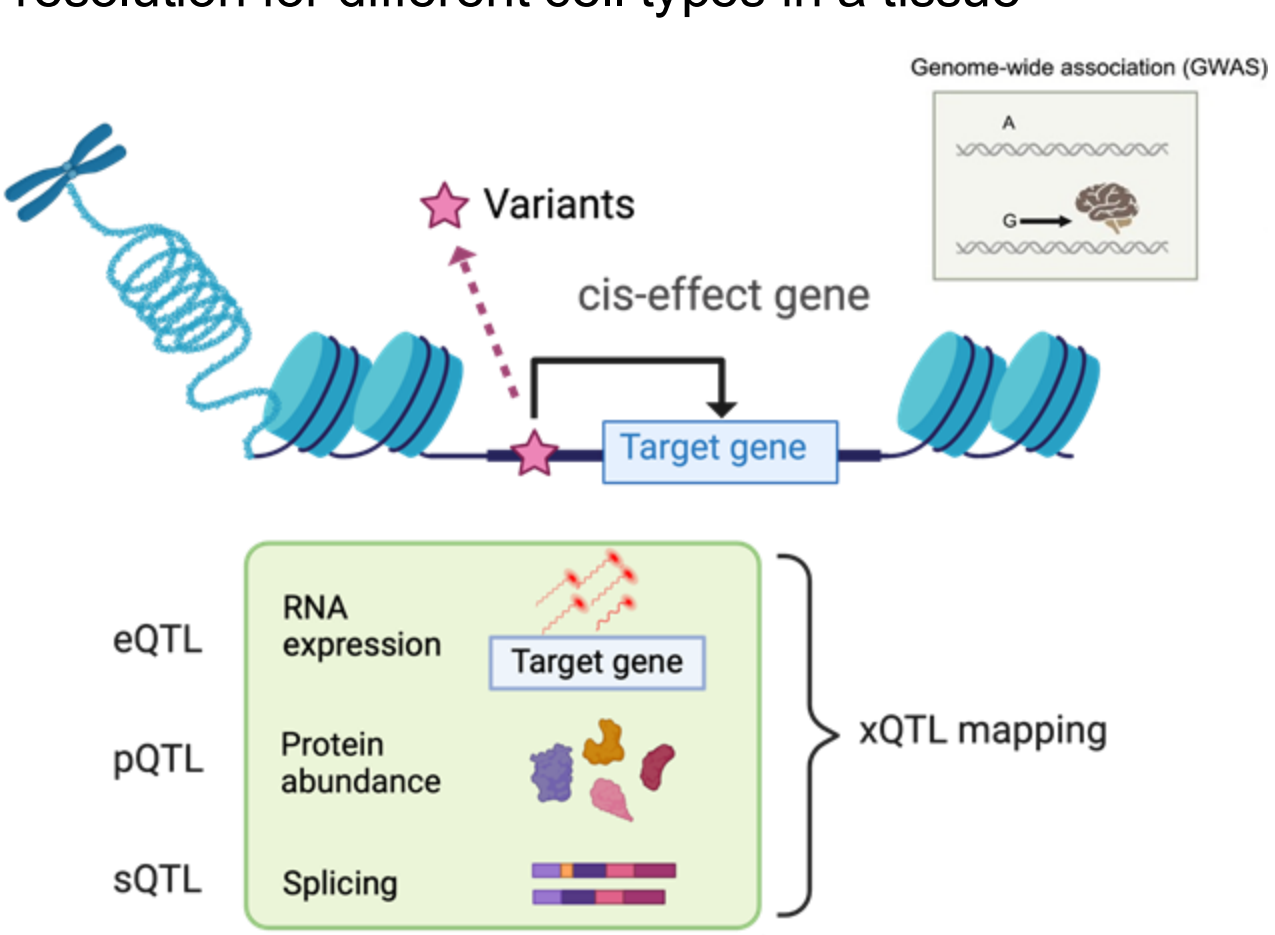


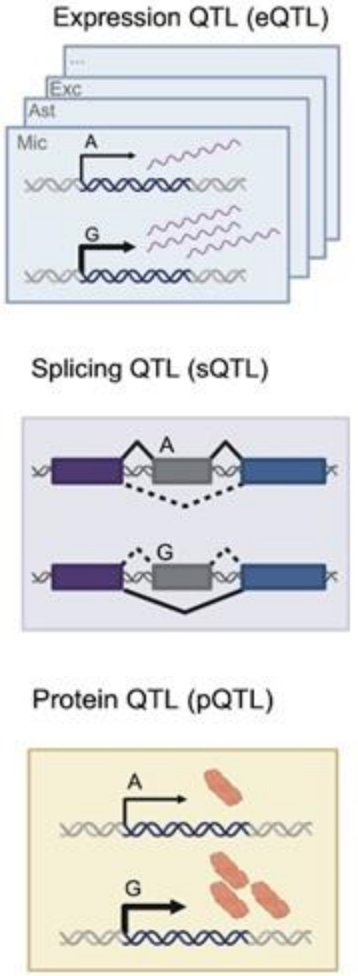
Coloc: standard method for colocalization.
Does not scale well to more than 2 phenotypes.

Giambartolomei et al 2014 *PLoS Gen*

ColocBoost model to perform multimodal molecular phenotype QTL colocalization

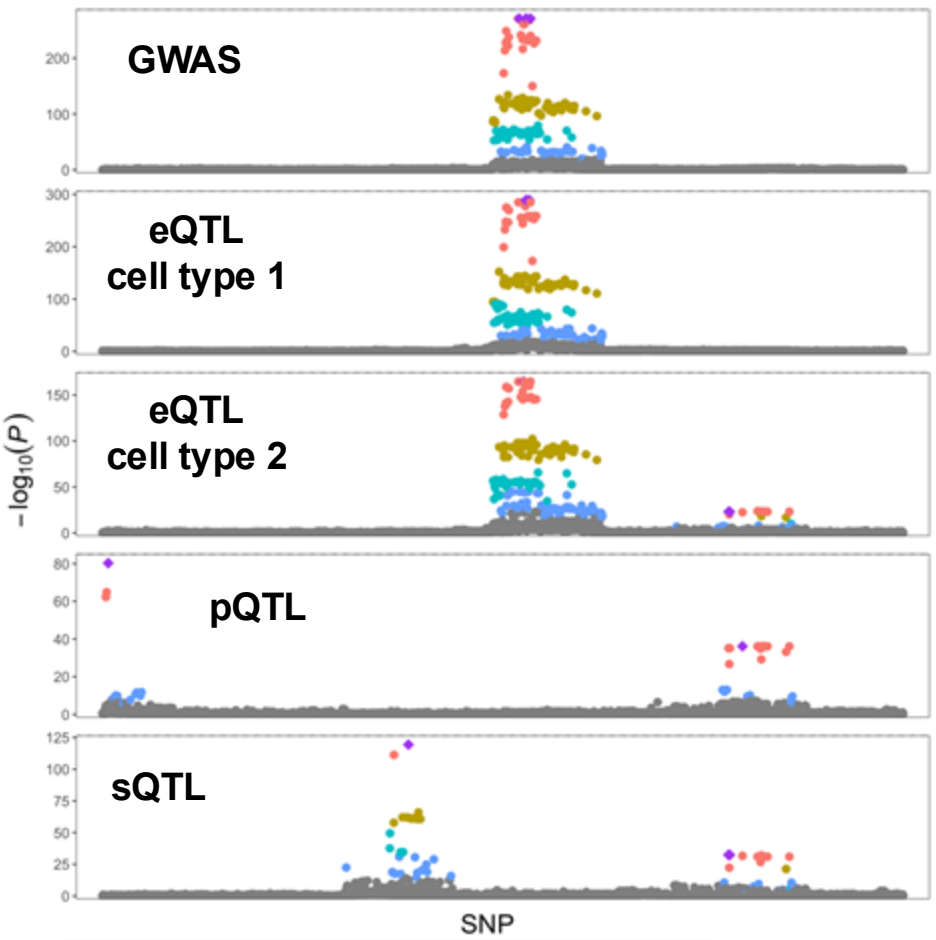
Recent advances in technology has made it easier to use other molecular phenotypes outside of gene expression, and also assess eQTL at cell type resolution for different cell types in a tissue

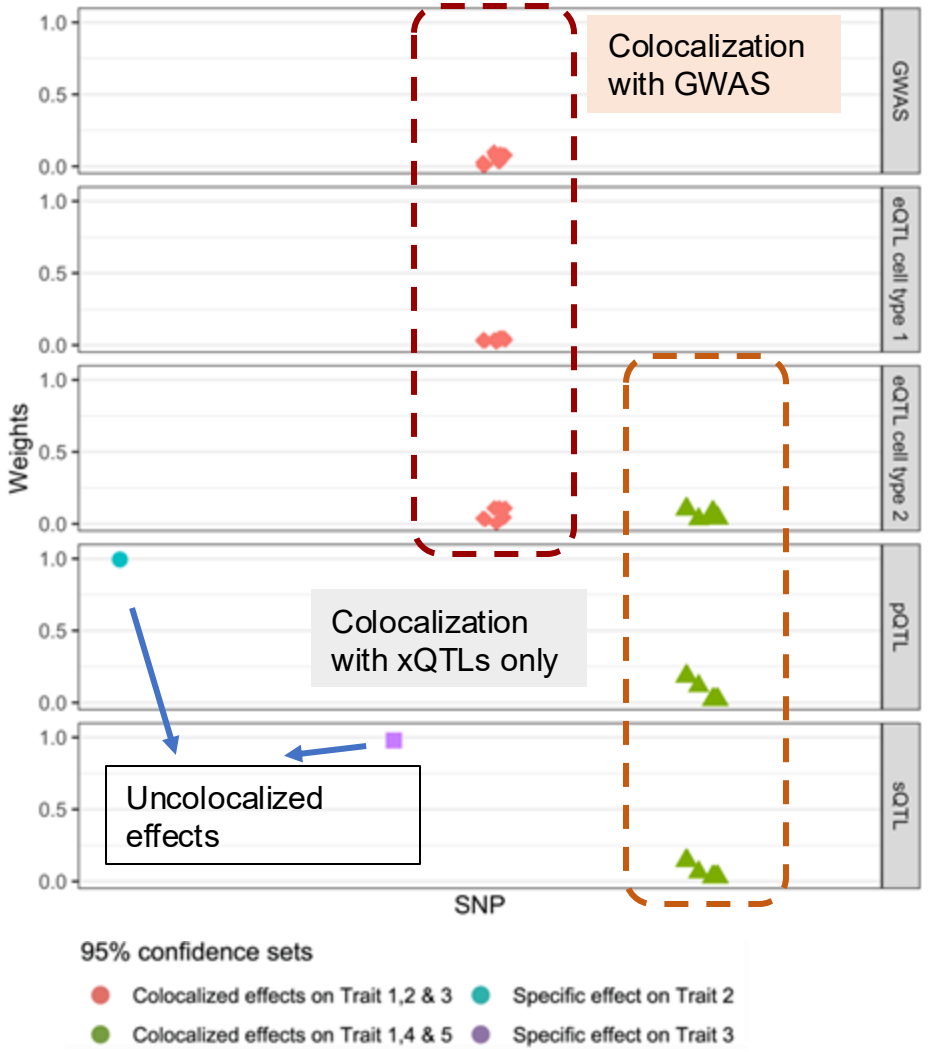




Aguet et al. 2020. *Nature Reviews Methods Primers*

Understanding colocalization: enhancing GWAS insights through shared genetic signals



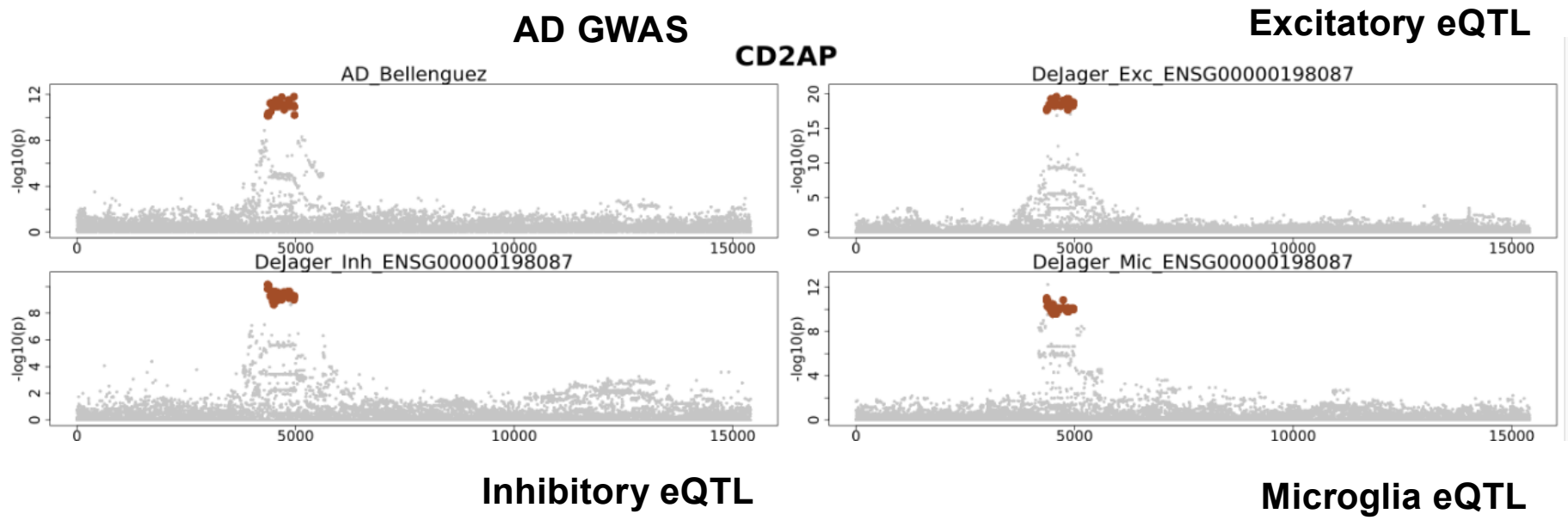


Memorial Sloan Kettering Cancer Center

ColocBoost: Cao et al 2025 medRxiv, in rev *Nat Genet*
HyPrColoc: Foley et al 2021 *Nat Commun*

Shared genetic regulation across cell types observed for many disease risk variants are not indicative of cell-cell crosstalk

37.3% of AD causal risk variants show genetic regulation shared across multiple cell types in brain.



Alzheimer's disease risk gene *CD2AP* is a dose-sensitive determinant of synaptic structure and plasticity

Matea Pavešković [1,2,3], Ruth B De-Paula [4,5,6], Shamsideen A Ojelade [7,8], Evelyne K Tantry [9,10], Mikhail Y Kochukov [11,12], Suyang Bao [13,14], Surabi Veeraragavan [15,16], Alexandra R Garza [17,18], Snigdha Srivastava [19,20,21], Si-Yuan Song [22,23], Masashi Fujita [24], Duc M Duong [25], David A Bennett [26], Philip L De Jager [27], Nicholas T Seyfried [28], Mary E Dickinson [29,30], Jason D Heaney [31], Benjamin R Arenkiel [32,33,34,#], Joshua M Shulman [35,36,37,38,39,#,✉]

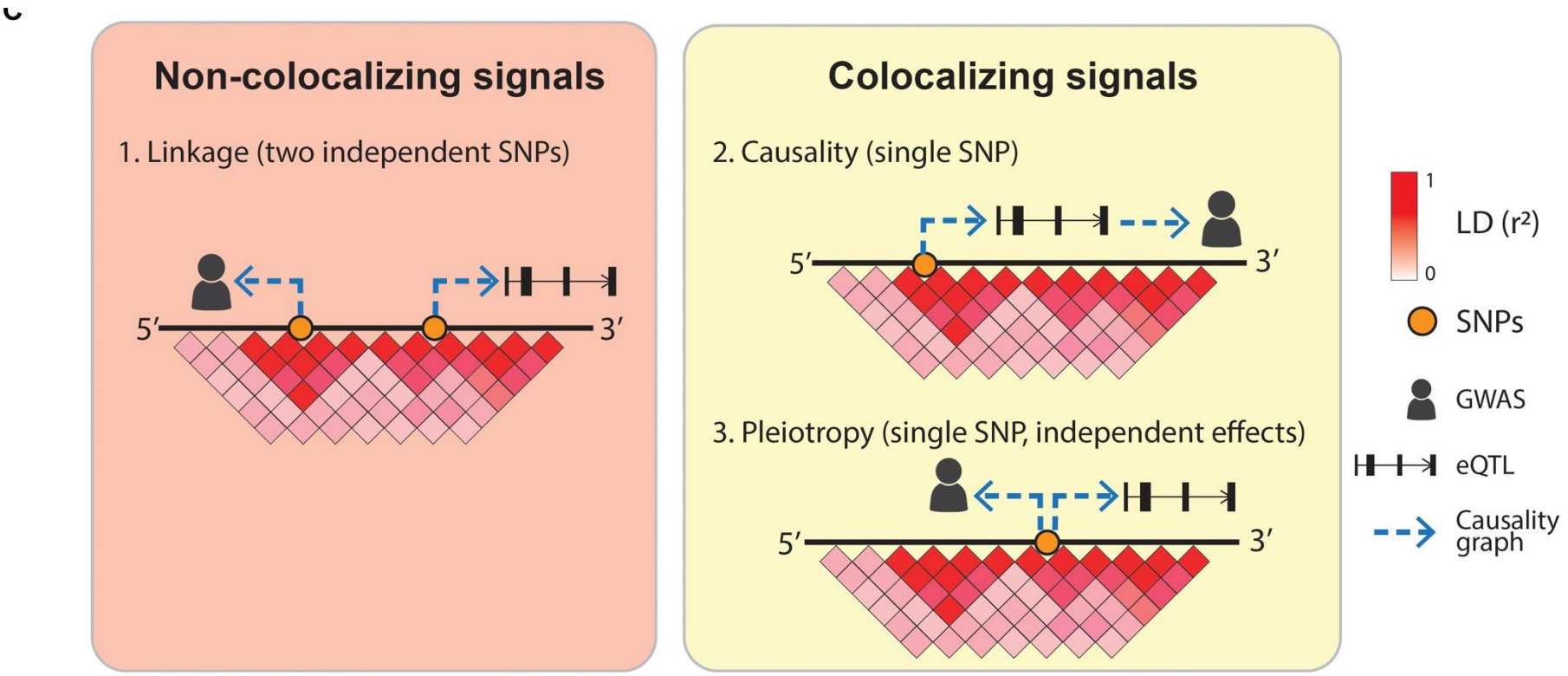
Microglial CD2AP deficiency exerts protection in an Alzheimer's disease model of amyloidosis

Lingliang Zhang [# 1], Lingling Huang [# 1], Yuhang Zhou [1], Jian Meng [1], Liang Zhang [1], Yunqiang Zhou [1], Naizhen Zheng [1], Tiantian Guo [1], Shanshan Zhao [1], Zijie Wang [1], Yuanhui Huo [1], Yingjun Zhao [1], Xiao-Fen Chen [1], Honghua Zheng [1], David M Holtzman [2], Yun-Wu Zhang [3]

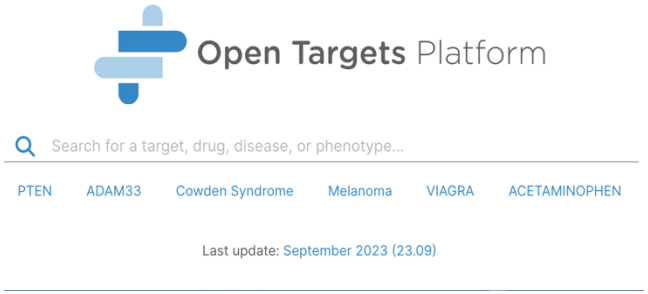
Paveskovic et al 2024 *HMG*
Zhang et al *2024 Mol Neurodeger.*

ColocBoost: Cao et al 2025 medRxiv, in rev *Nat Genet*
HyPrColoc: Foley et al 2021 *Nat Commun*

Colocalization of GWAS and eQTL signals to identify shared causal variants between disease and expression



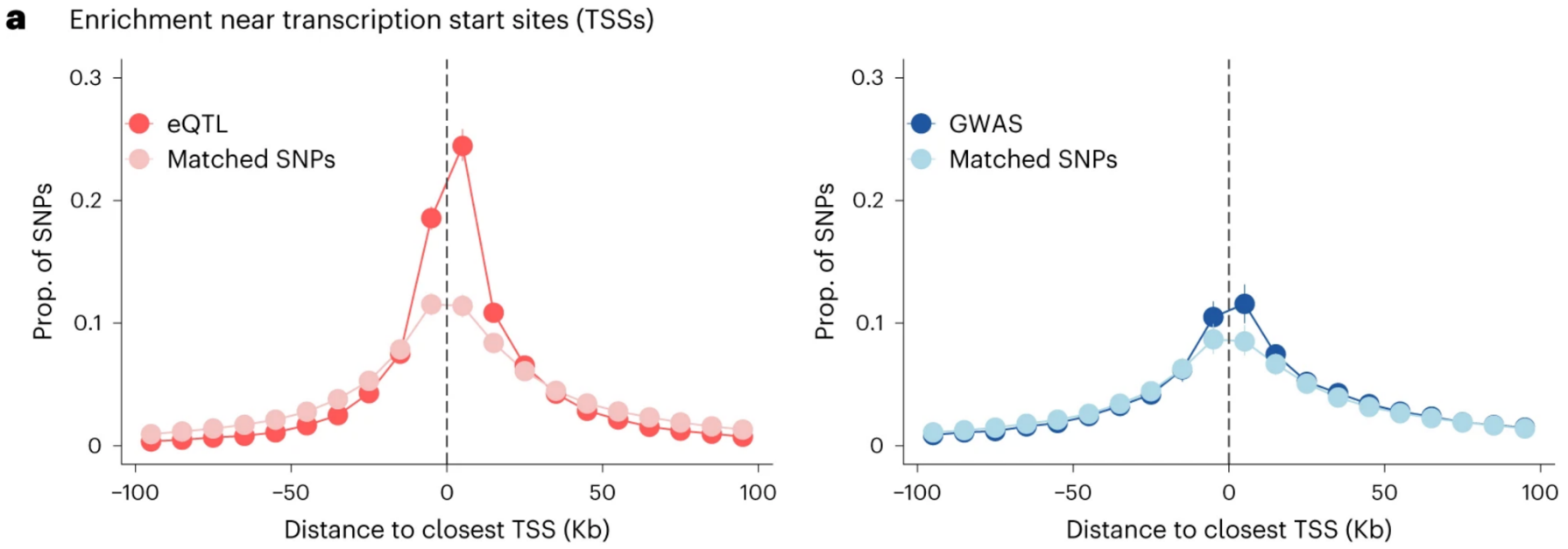
Integrating eQTL and enhancer-gene maps to link variants to genes (OpenTargets)



Data type	Experiment type	Source	Weighting
In silico functional prediction	Transcript consequence	VEP	1.0
QTL	sQTL	GTEx v8	1.0
QTL	eQTL	*many*	0.66
QTL	pQTL	*many*	0.66
Interaction	PCHi-C	Javierre *et al.* (Cell, 2016)	0.33
Interaction	Enhancer-TSS correlation	Andersson *et al.* (Nature, 2014)	0.33
Interaction	DHS-promoter correlation	Thurman *et al.* (Nature, 2012)	0.33
Distance	Canonical TSS		0.33

Mountjoy et al 2021 *Nat Genet*

Systemic differences between eQTLs and GWAS



› Nat Genet. 2023 Nov;55(11):1866-1875. doi: 10.1038/s41588-023-01529-1. Epub 2023 Oct 19.

Systematic differences in discovery of genetic effects on gene expression and complex traits

Hakhamanesh Mostafavi [1], Jeffrey P Spence [2], Sahin Naqvi [2 3], Jonathan K Pritchard [4 5]

Affiliations + expand

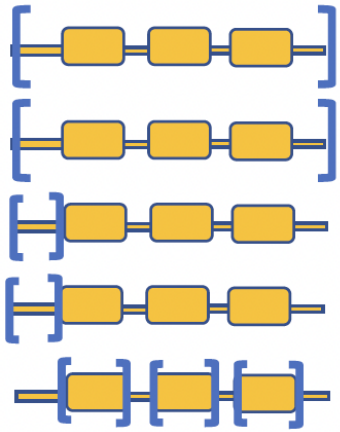
PMID: 37857933 DOI: 10.1038/s41588-023-01529-1

"GWAS and cis-eQTL hits are systematically different: ***eQTLs cluster strongly near transcription start sites, whereas GWAS hits do not. Genes near GWAS hits are enriched in key functional annotations****, are under strong selective constraint and have complex regulatory landscapes across different tissue/cell types, whereas genes near* ***eQTLs are depleted of most functional annotations****.*

Mostafavi et al 2023 *Nat Genet*

Linking GWAS variants to target genes

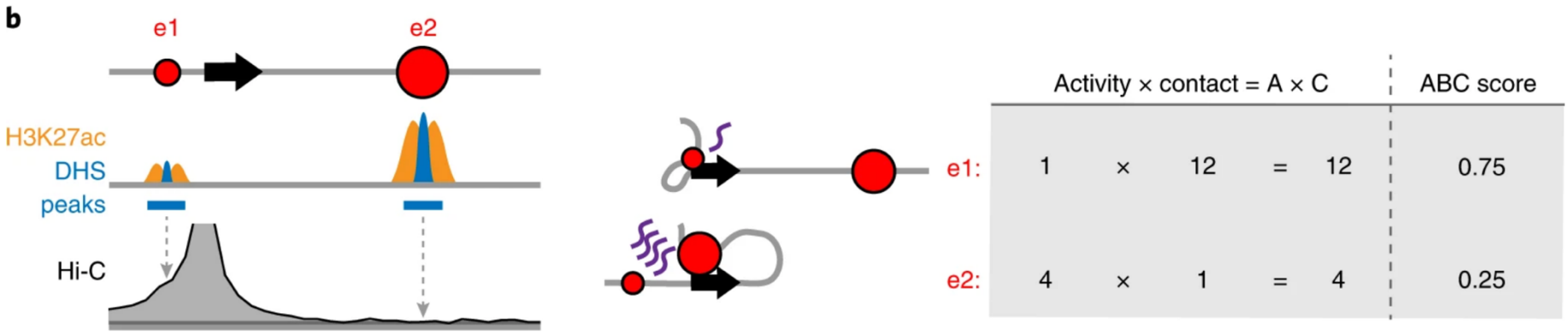
Broadening the scope of approaches to link variants to genes



S2G strategies	Description
5kb	SNPs in 5kb window around gene
100kb	SNPs in 100 kb window around gene
Promoter	SNPs in promoter region of the gene
TSS	SNPs in and around Transcription start sites
Coding	SNPs in coding regions of the gene

Dey et al 2022, *Cell Genomics,*
Gazal..Dey et al 2022 *Nat Genet*

The Activity-By-Contact element-gene linking method

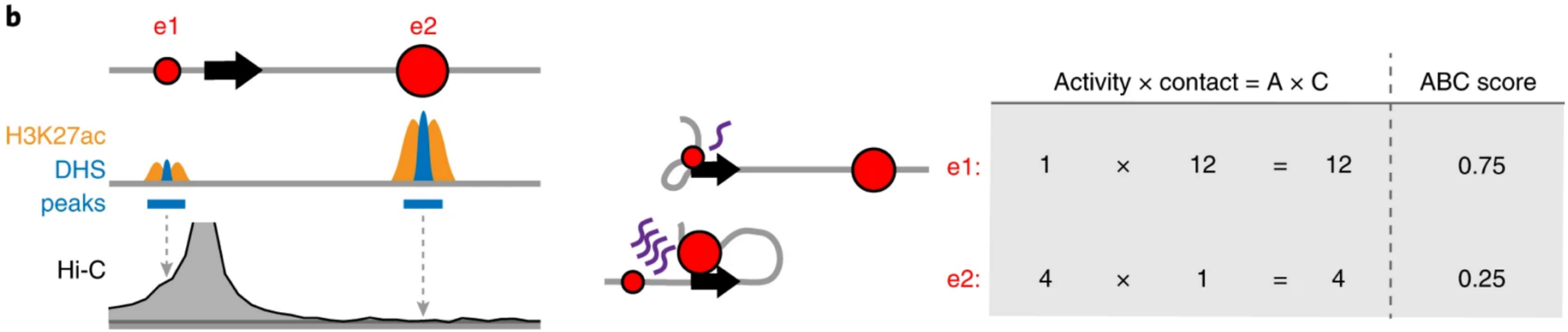


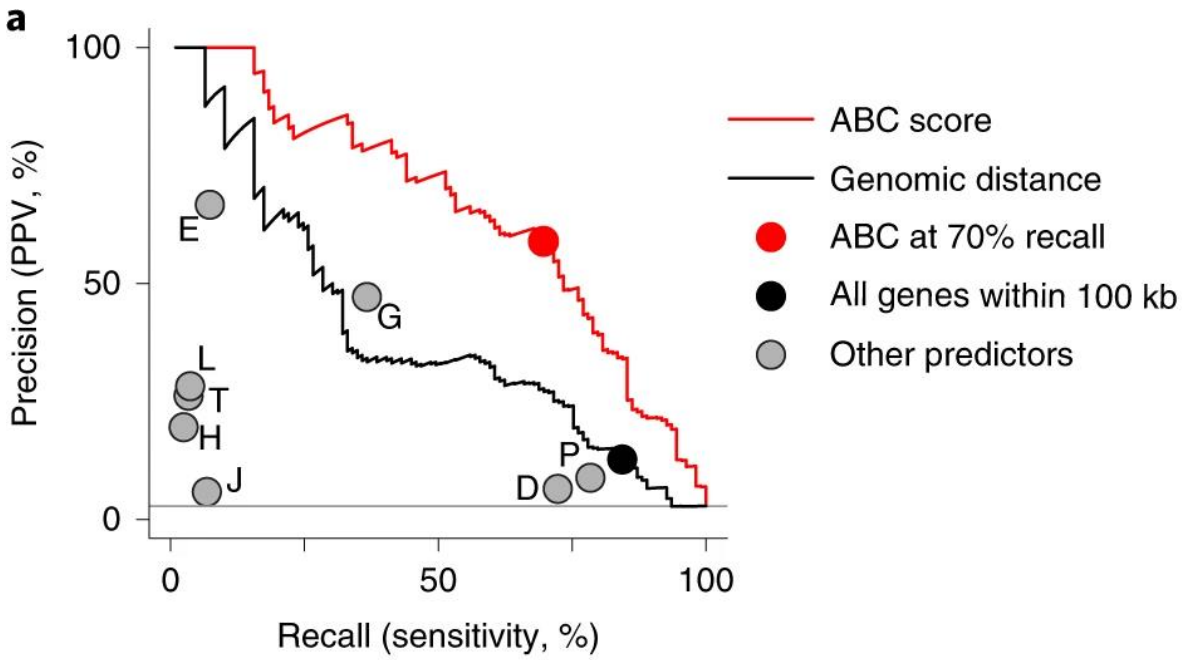
$$\text{ABC score}_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{\text{all elements e within 5 Mb of } G} A_e \times C_{e,G}}$$

Operationally, we estimated Activity (A) as the geometric mean of the read counts of DHS and H3K27ac chromatin immunoprecipitation sequencing (ChIP–seq) at element E, and Contact (C) as the KR-normalized Hi-C contact frequency between E and the promoter of gene G at 5-kb resolution (see Supplementary Note 4 and Supplementary Figs. 4 and 5).

Fulco et al 2019 *Nat Genet*
Nasser et al 2021 *Nature*

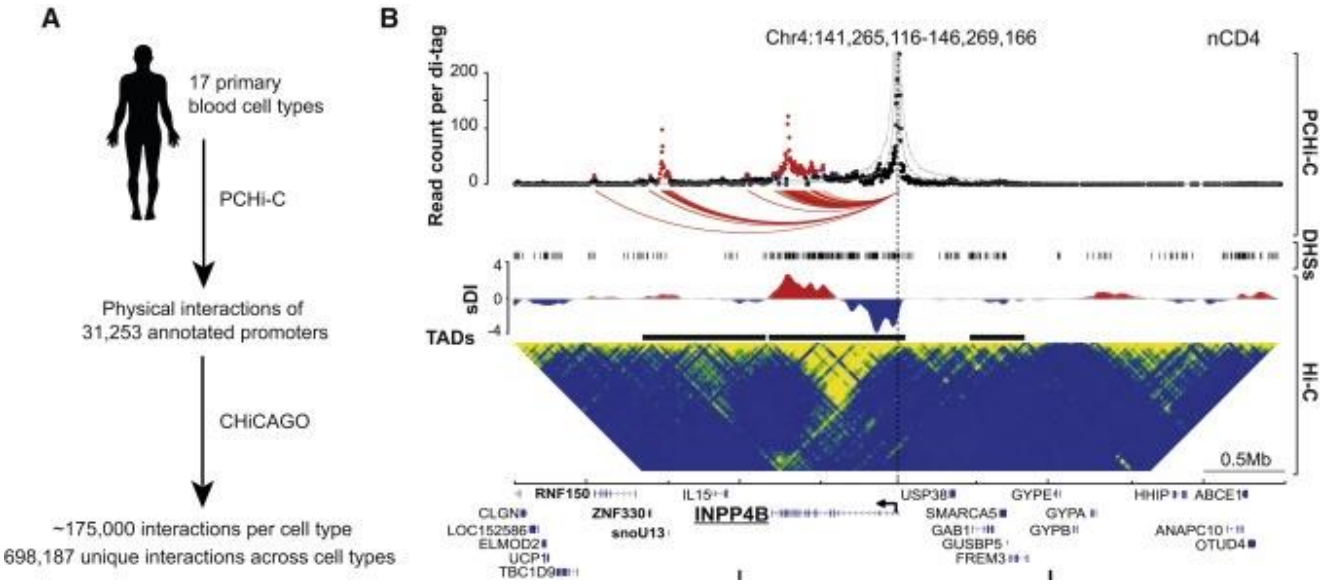
The Activity-By-Contact element-gene linking method



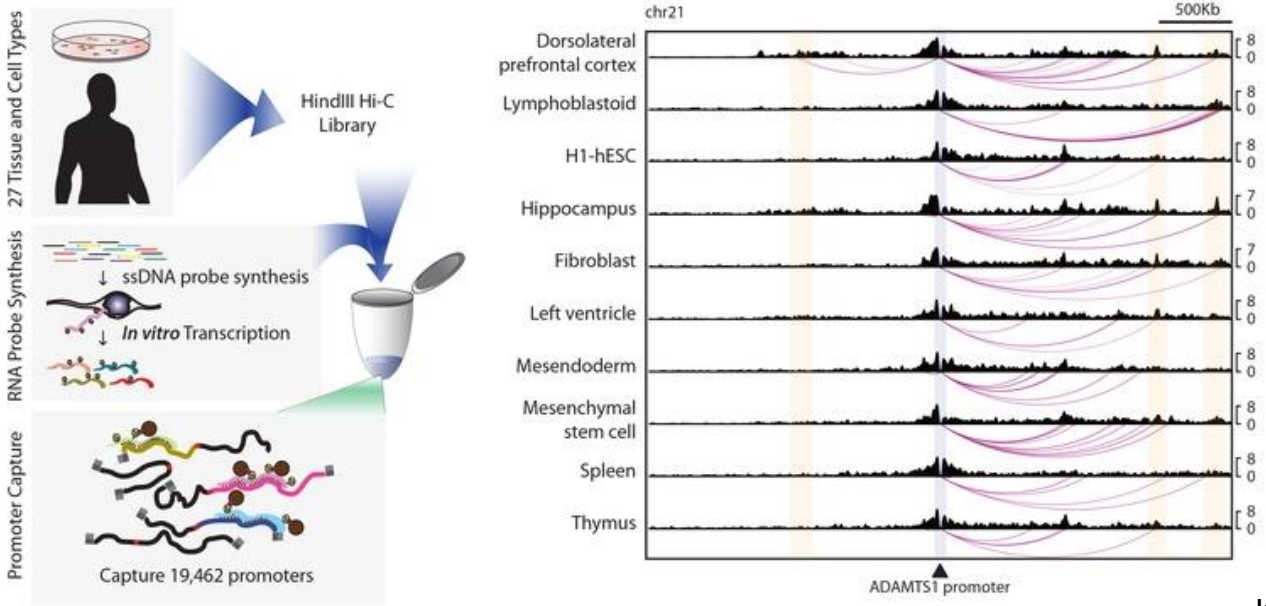


Fulco et al 2019 *Nat Genet*
Nasser et al 2021 *Nature*

Promoter-capture Hi-C to link elements to genes



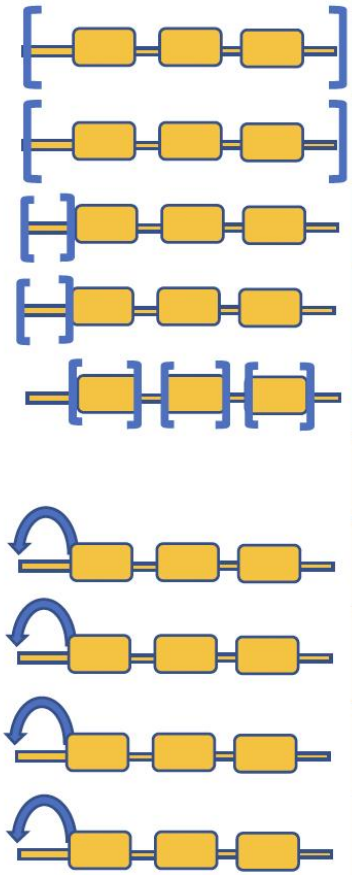
Javierre et al 2016 Cell



Jung et al 2020 *Nat Genet*

Broadening the scope of approaches to link variants to genes

Naive S2G | Expression S2G | Hi-C S2G

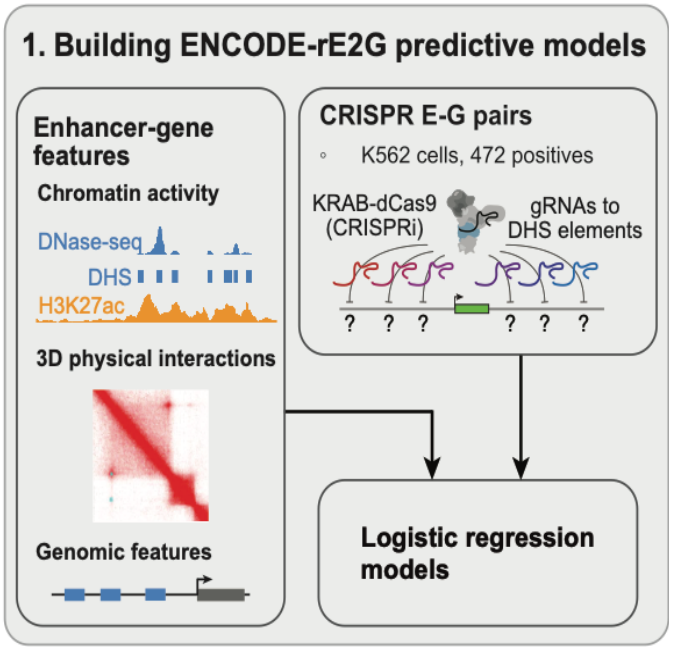


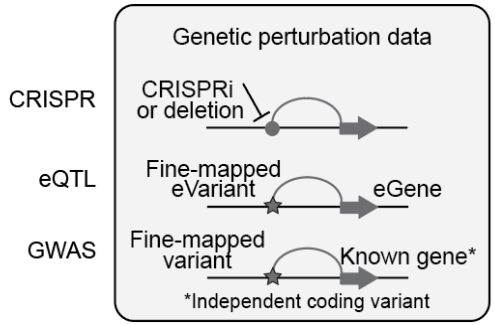
S2G strategies	Description
5kb	SNPs in 5kb window around gene
100kb	SNPs in 100 kb window around gene
Promoter	SNPs in promoter region of the gene
TSS	SNPs in and around Transcription start sites
Coding	SNPs in coding regions of the gene
eQTL	Max. post. causal probability in GTEx blood[1,2]
ATAC	Correlated ATAC-seq peaks and gene expression in blood[3]
Roadmap	Correlated enhancers and gene expression in blood[4,5,6]
PC-HiC	Promoter Capture Hi-C[7]
ABC	DHS ∩ H3K27ac ∩ Hi-C in blood[8]

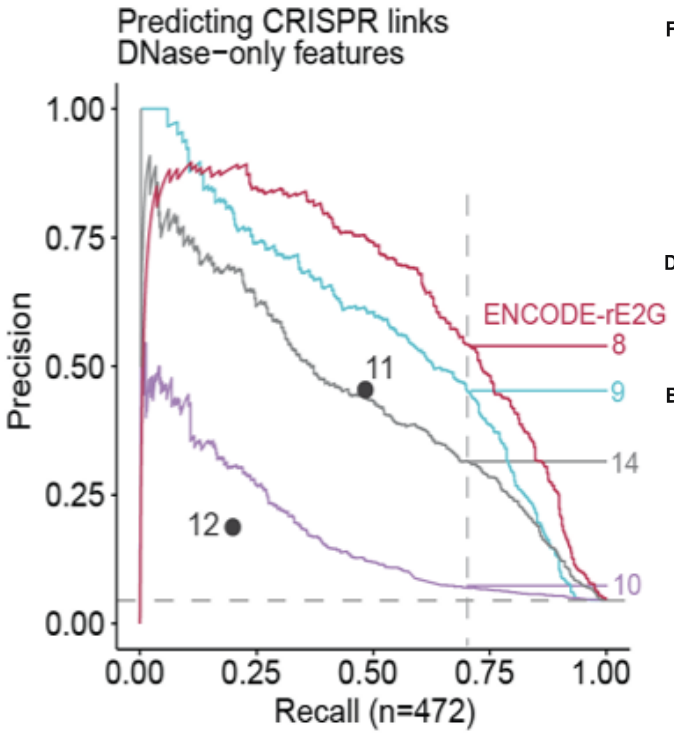
[1]Hormozdiari et al 2018 NG, [2]Aguet et al 2019 bioRxiv, [3]Yoshida et al 2019 Cell, [4]Liu et al 2017 Gen. Biol. , [5]Ernst et al 2011 Nat. meth., [6]Kundaje et al 2015 Nature , [7]Javierre et al 2016 Cell,, [8]Fulco et al 2019 Nat.Genet

Dey et al 2022, *Cell Genomics,*
Gazal..Dey et al 2022 *Nat Genet*

Benchmarking different element-gene linking approaches







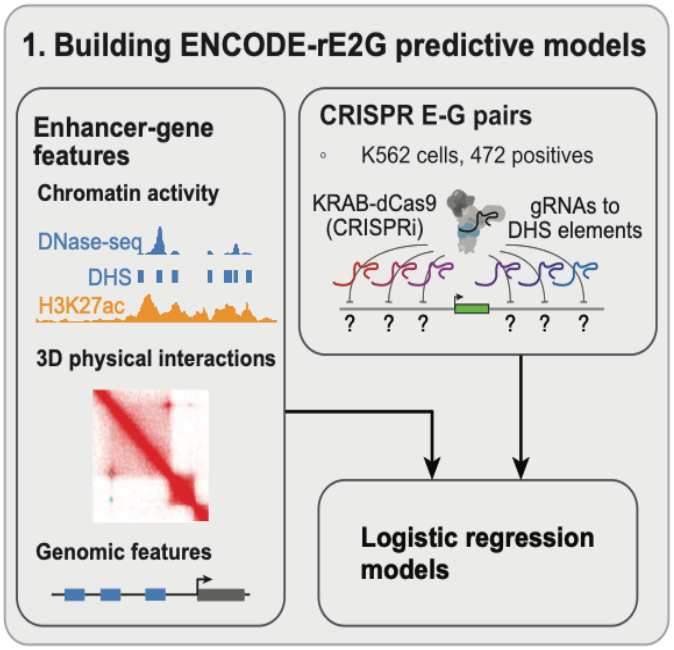
Full features
- (1) ENCODE−rE2G^{Extended}
- (2) ABC$^{A=DNase\ x\ H3K27ac,\ C=ENCODE\ Hi-C}$
- (3) CIA
- (4) EPIraction
- (5) Enformer
- (6) GraphReg
- (7) EpiMap

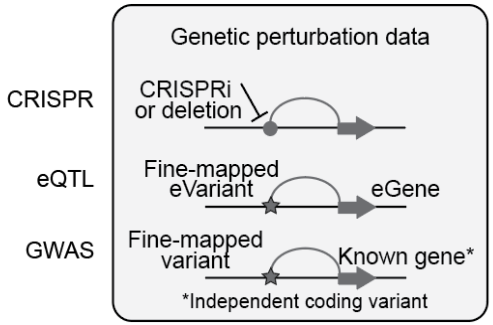
DNase-seq only
- (8) ENCODE−rE2G
- (9) ABC$^{A=DNase,\ C=Average\ ENCODE\ Hi-C}$
- (10) Correlation$^{E-P\ DNase-seq\ signal}$

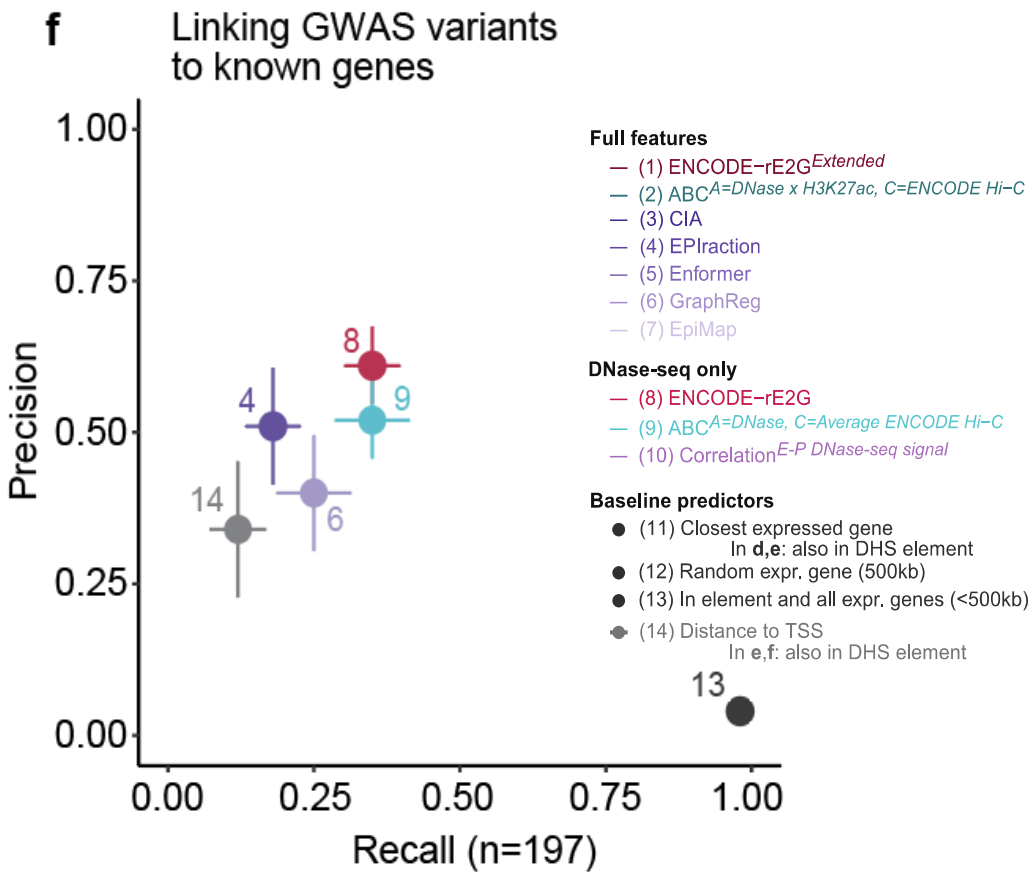
Baseline predictors
- (11) Closest expressed gene
  In **d**,**e**: also in DHS element
- (12) Random expr. gene (500kb)
- (13) In element and all expr. genes (<500kb)
- (14) Distance to TSS
  In **e**,**f**: also in DHS element

Gschwind*..,Dey*..Engreitz 2023 bioRxiv, in rev, *Nature*

Benchmarking different element-gene linking approaches

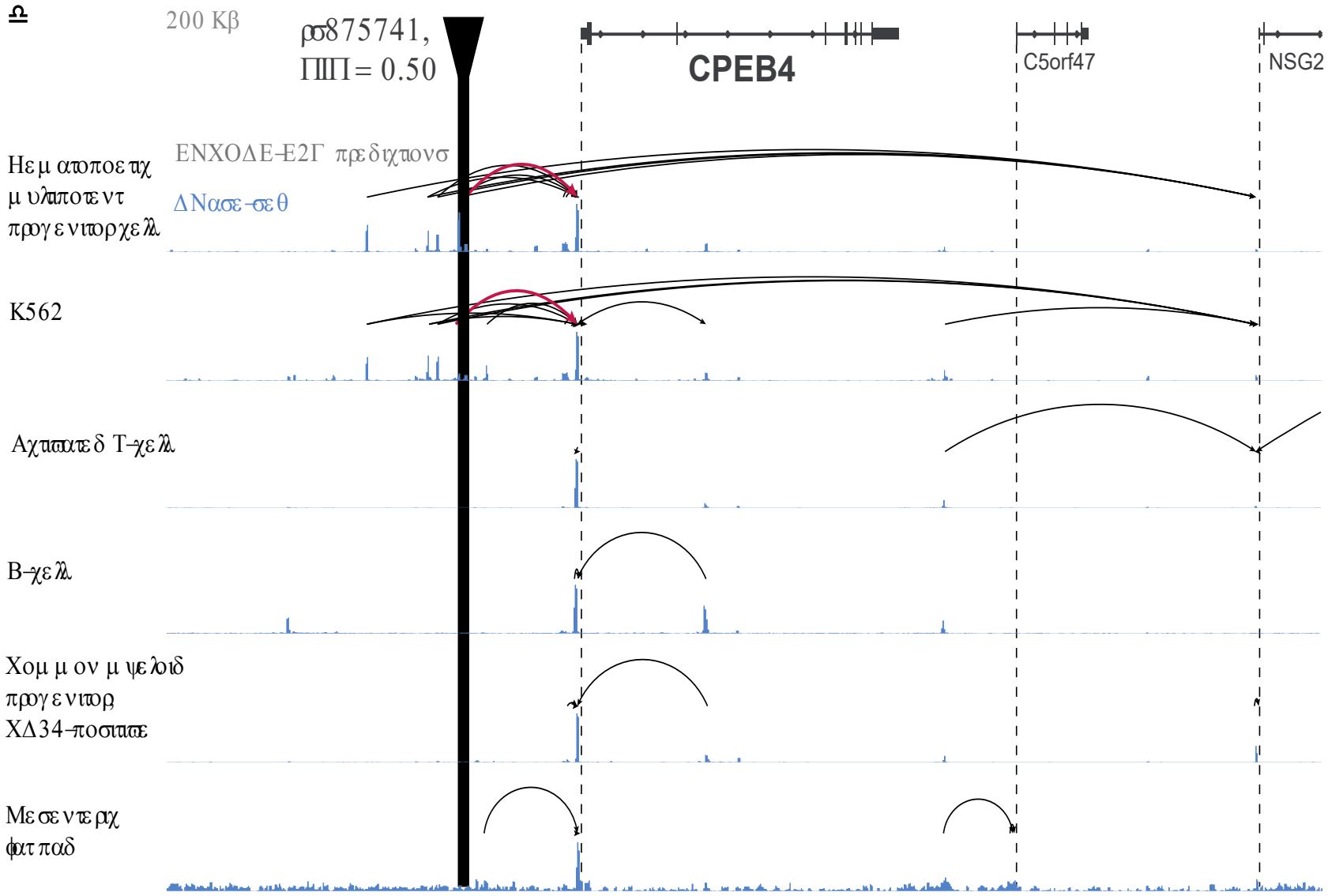




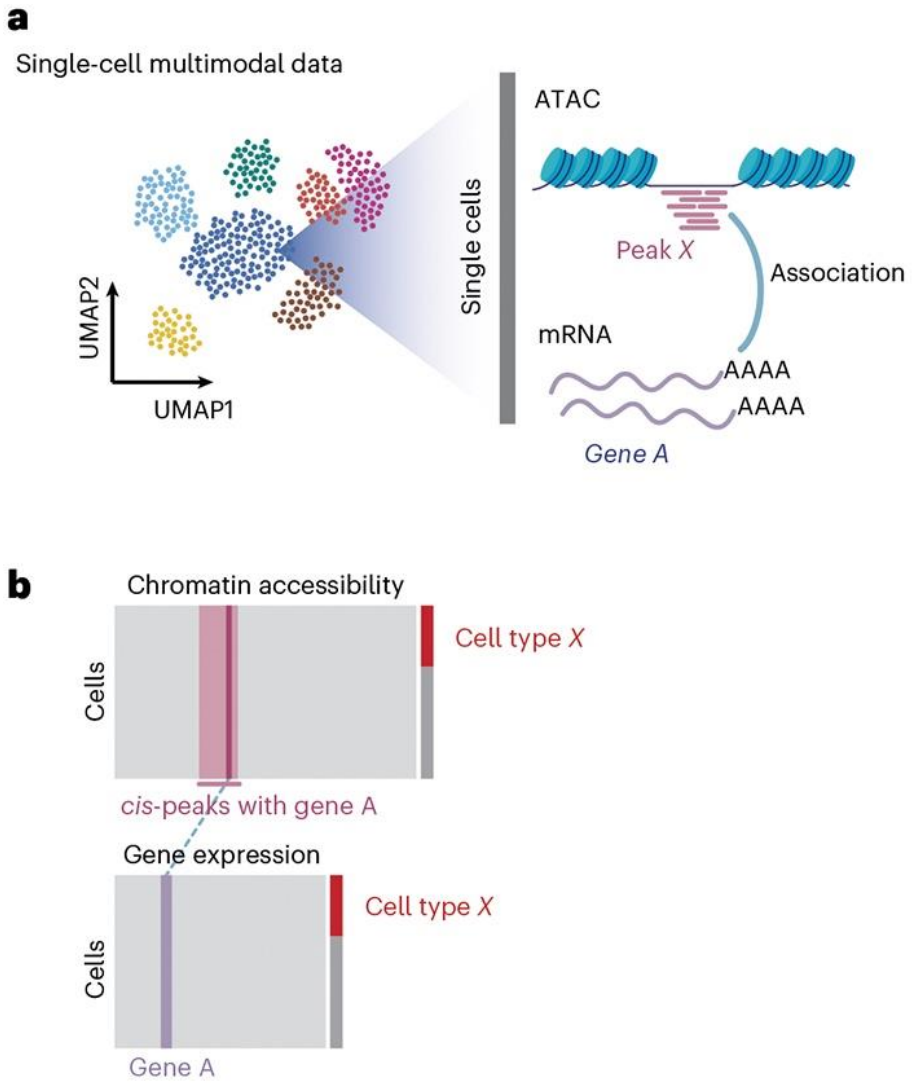


Gschwind*..,Dey*..Engreitz 2023 bioRxiv, in rev, *Nature*

Visualizing the element-gene links underlying **rs875741**: fine-mapped variant **PIP = 0.50 for mean corpuscular hemoglobin.**

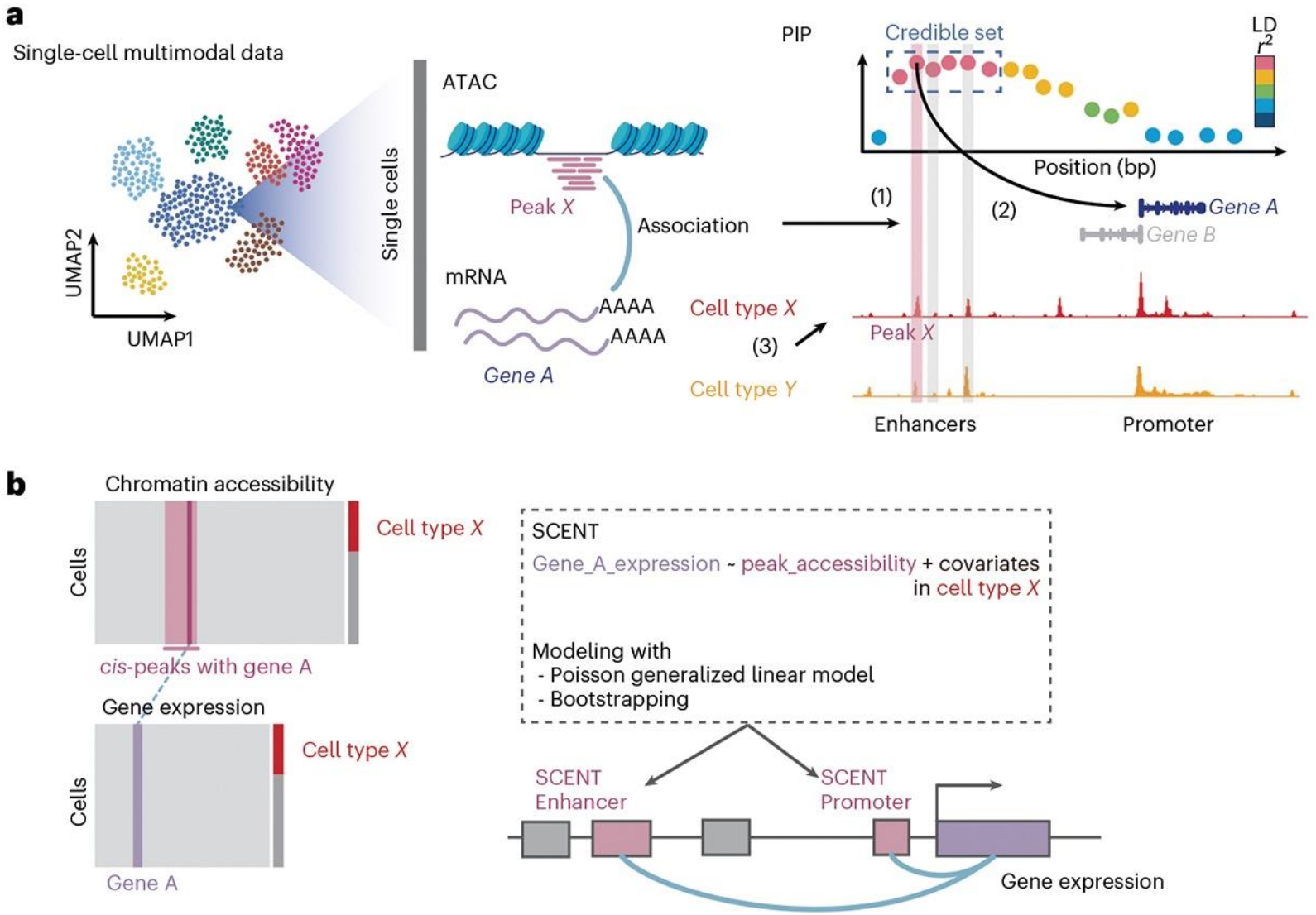


Using single-cell RNA and ATAC data in the same cell to identify regulatory elements linked to genes



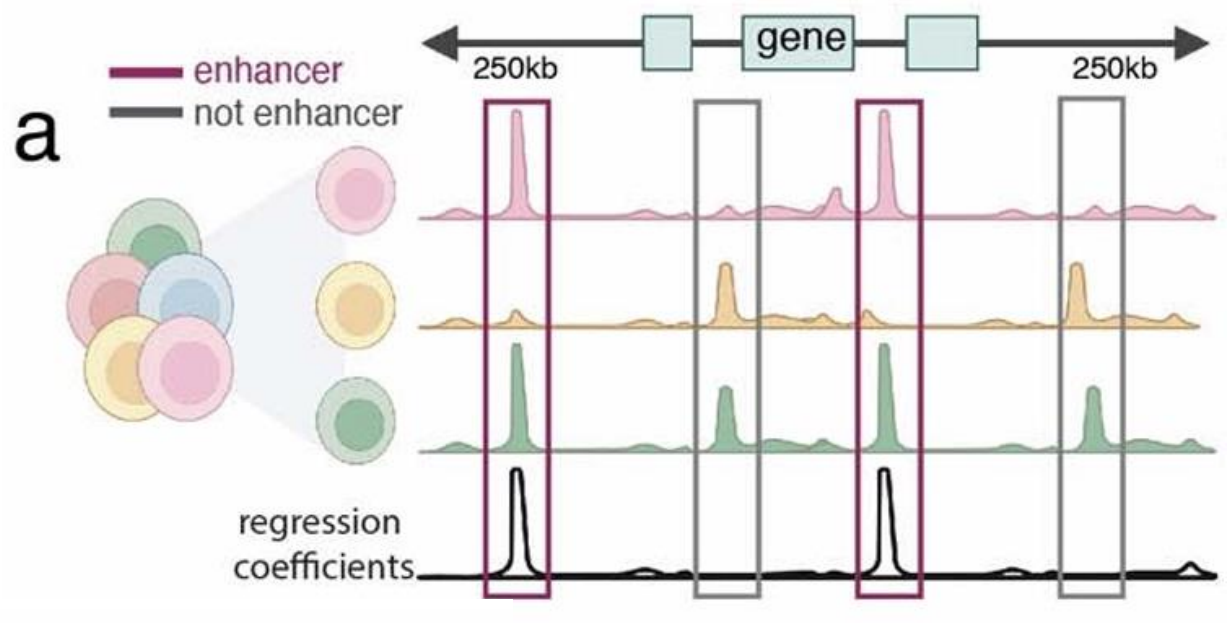
SCENT: Sakaue..Dey et al 2024 *Nat Genet*

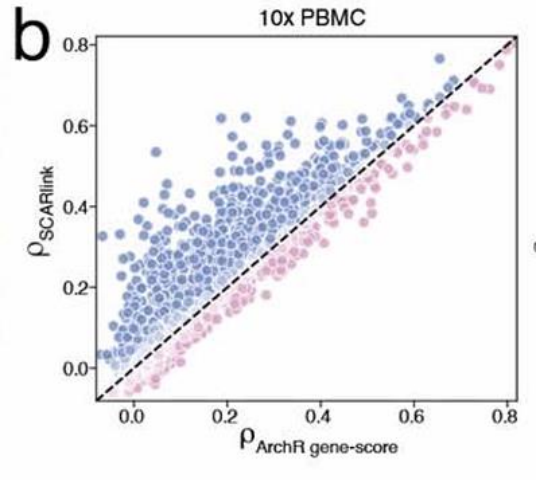
Using single-cell RNA and ATAC data in the same cell to identify regulatory elements linked to genes

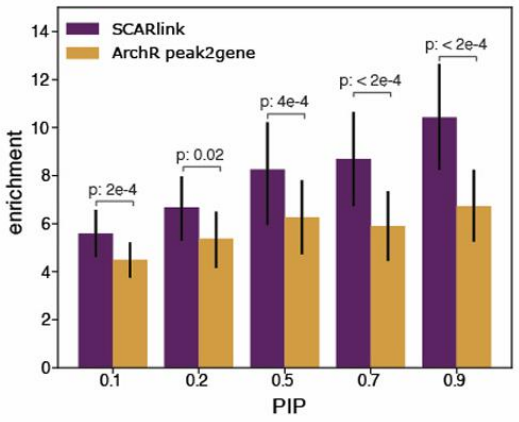


SCENT: Sakaue..Dey et al 2024 *Nat Genet*

Using single-cell RNA and ATAC data in the same cell to identify regulatory elements linked to genes

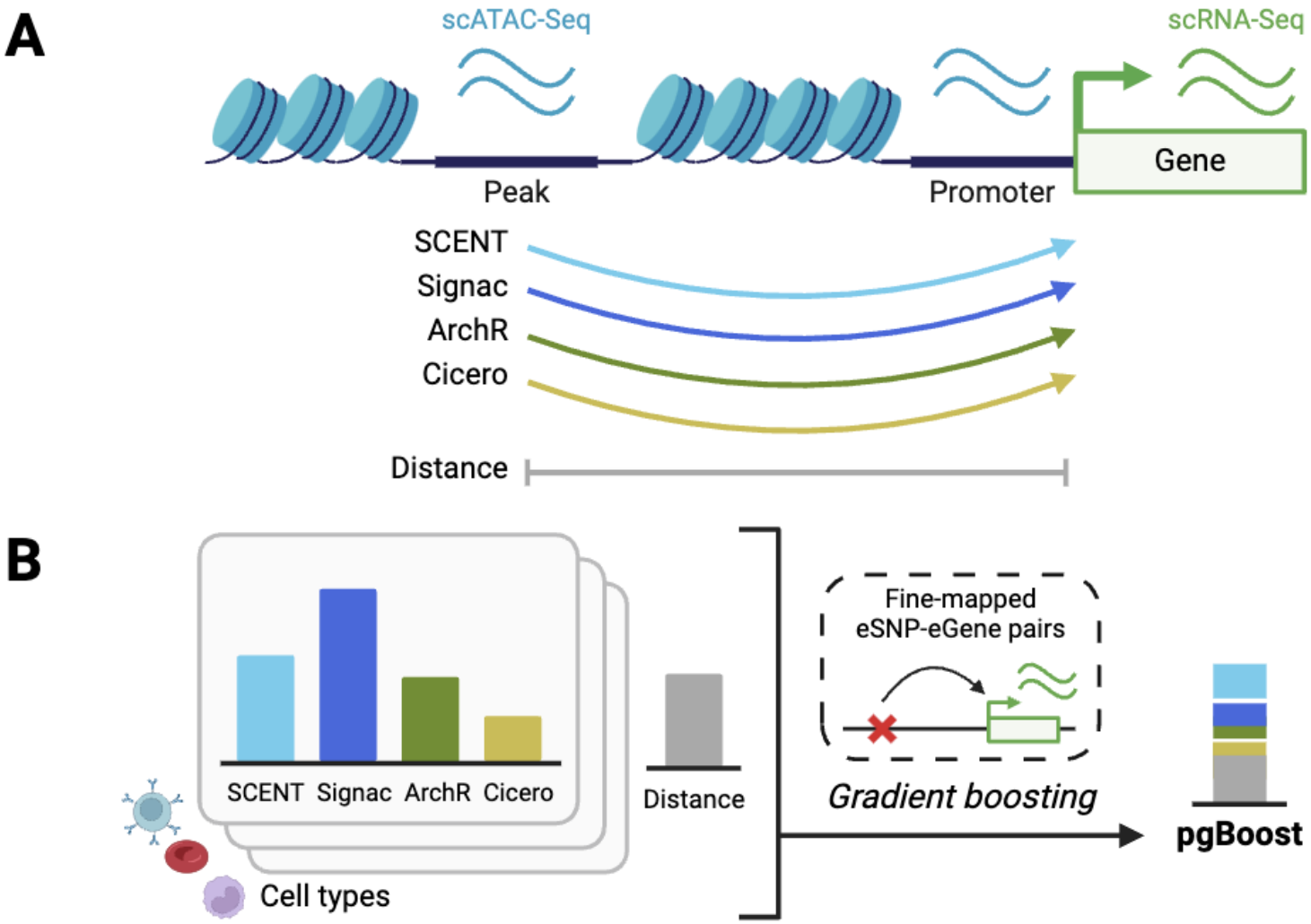






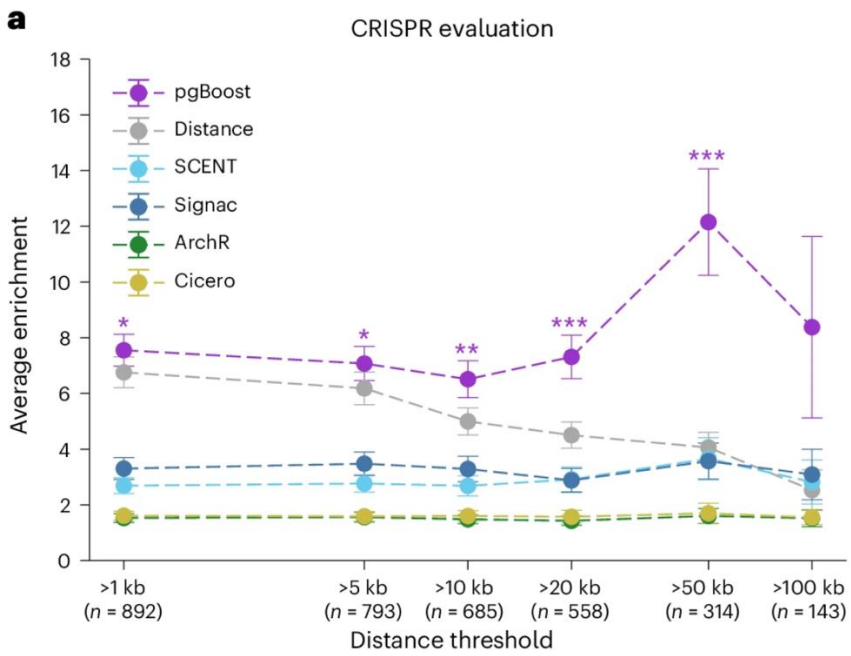
SCARlink: Mitra, .., Dey*, Leslie* 2024 *Nat Genet*

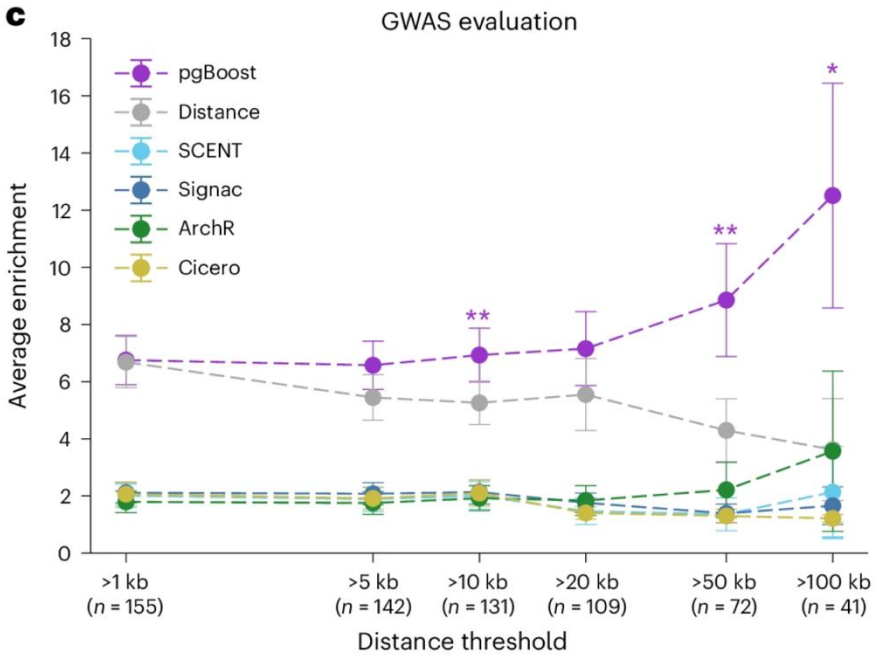
Integrative model of multiome peak-gene links and genomic distance is informative for diseases



pgBoost: Dorans, .., Dey*, Price* 2025 *Nat Genet*

Integrative model of multiome peak-gene links and genomic distance is informative for diseases

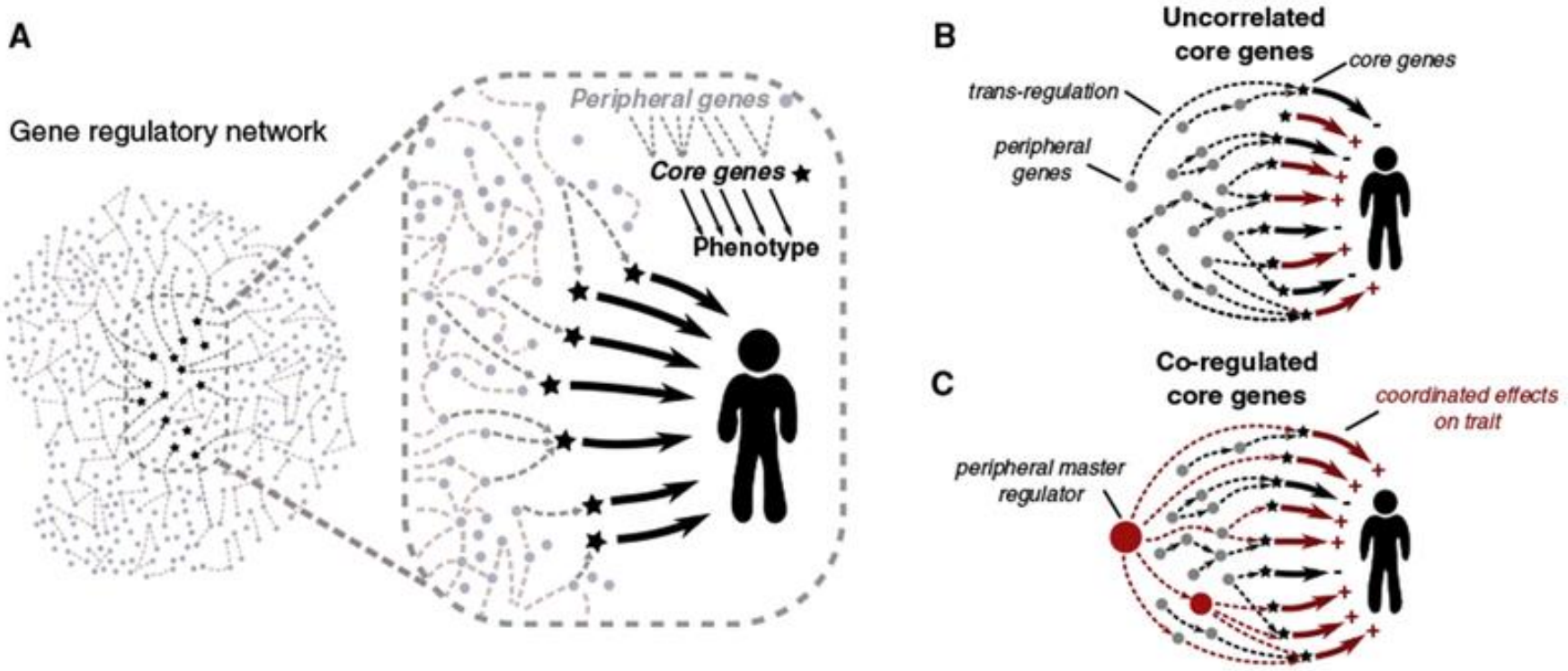




pgBoost: Dorans, .., Dey*, Price* 2025 *Nat Genet*

Linking genes and gene programs to GWAS disease risk

Beyond immediate target genes for GWAS variants, what are the gene programs or pathways affected by GWAS variants



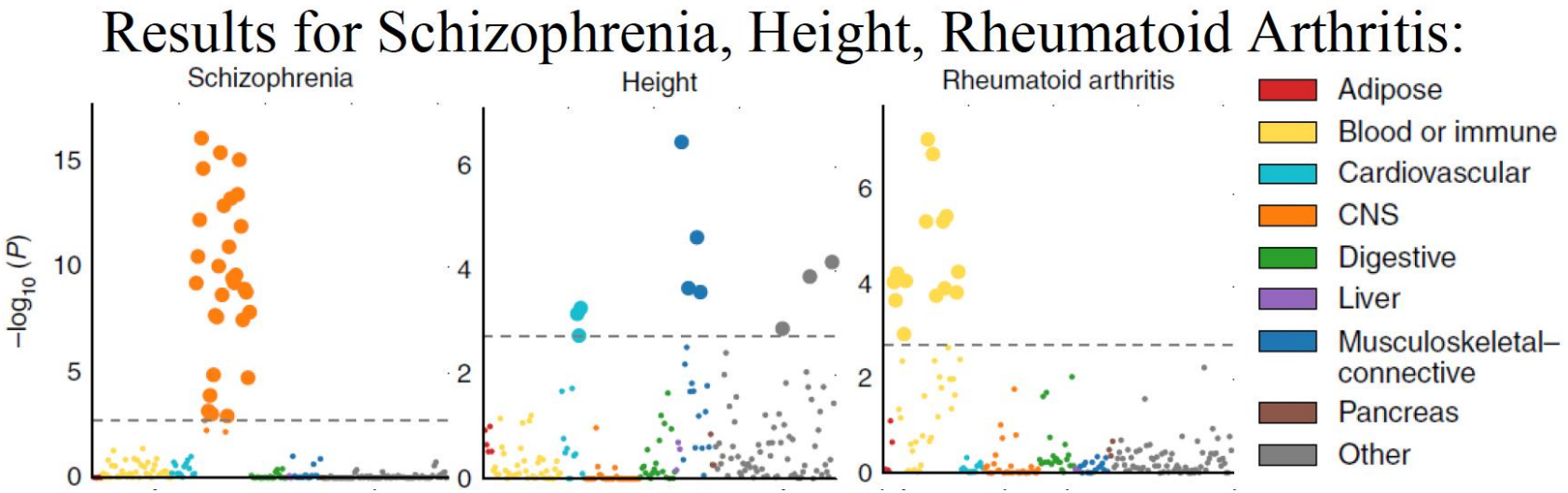
Liu, Li, Pritchard 2019 *Cell*

Mathieson 2021 *AJHG*
Boyle et al 2018 *Cell*

Specifically expressed genes in a tissue show tissue-specific heritability enrichments

SEG genes: Genes in top 10% of most enriched expression in a focal tissue compared to other GTEx tissues.

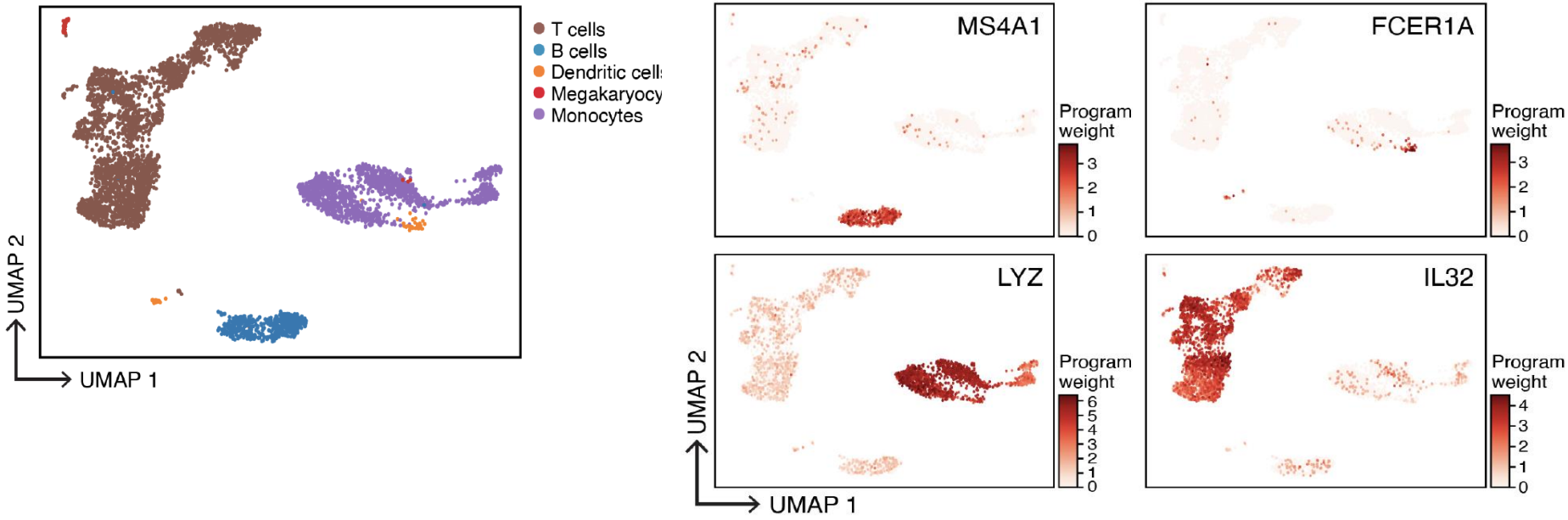
Annotate *variants if they lie in a 100KB window around the SEG genes*



Finucane et al 2018, *Nat Genet*

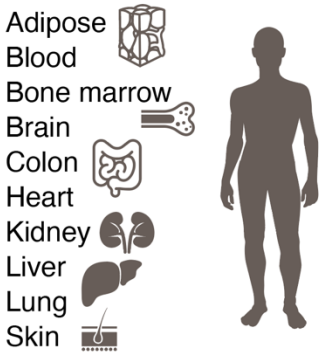
Constructing gene programs to characterize cell type

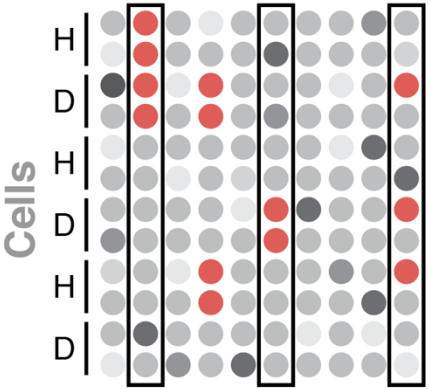
Cell type program: Genes specifically expressed in an annotated cell type compared to other cell types in the tissue [nonparametric DE: Wilcoxon rank-sum test].



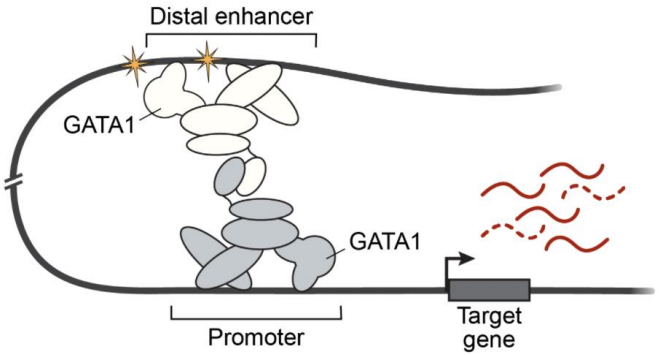
Jagadeesh*, Dey* et al 2022, *Nat Genet*

sc-linker: Using single-cell RNA-seq to assess cell-type-specific heritability enrichments of specifically expressed genes





Step 1: Gene programs



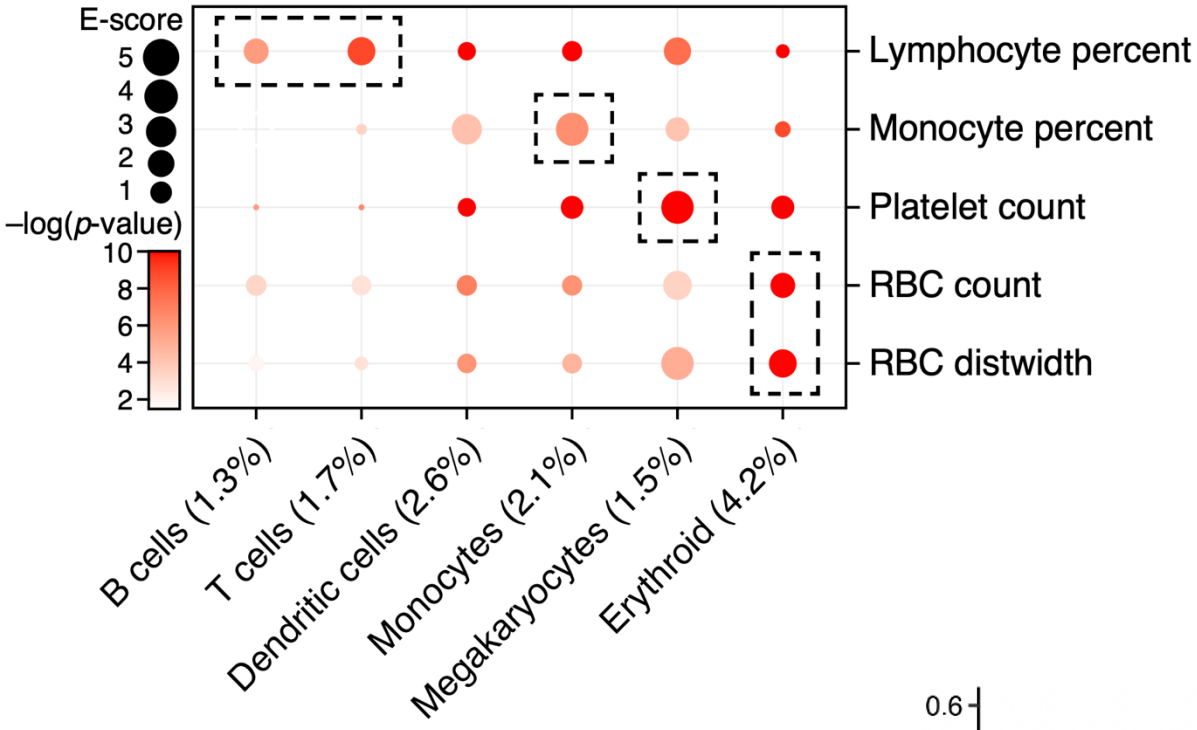
Step 2: SNP-gene maps to generate SNP annotation from gene program

Activity-By-Contact (ABC)
U Roadmap enhancer-gene

Step 3: Disease heritability enrichment

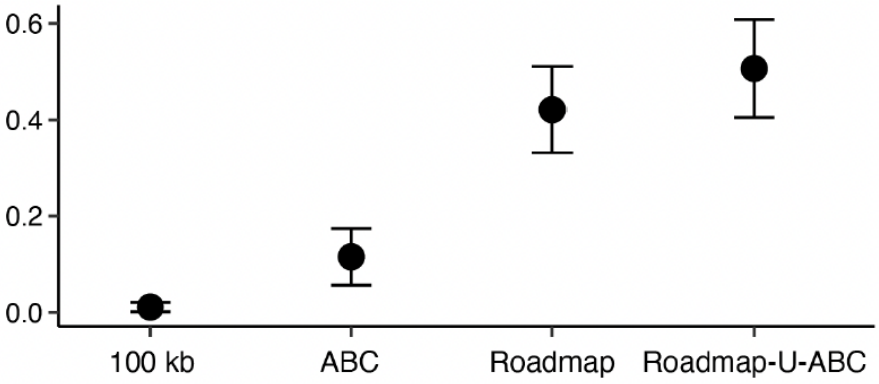
Jagadeesh*, Dey* et al 2022 *Nat Genet*;
Delorey* … Dey* et al 2021, *Nature*

sc-linker: Using single-cell RNA-seq to assess cell-type-specific heritability enrichments of specifically expressed genes

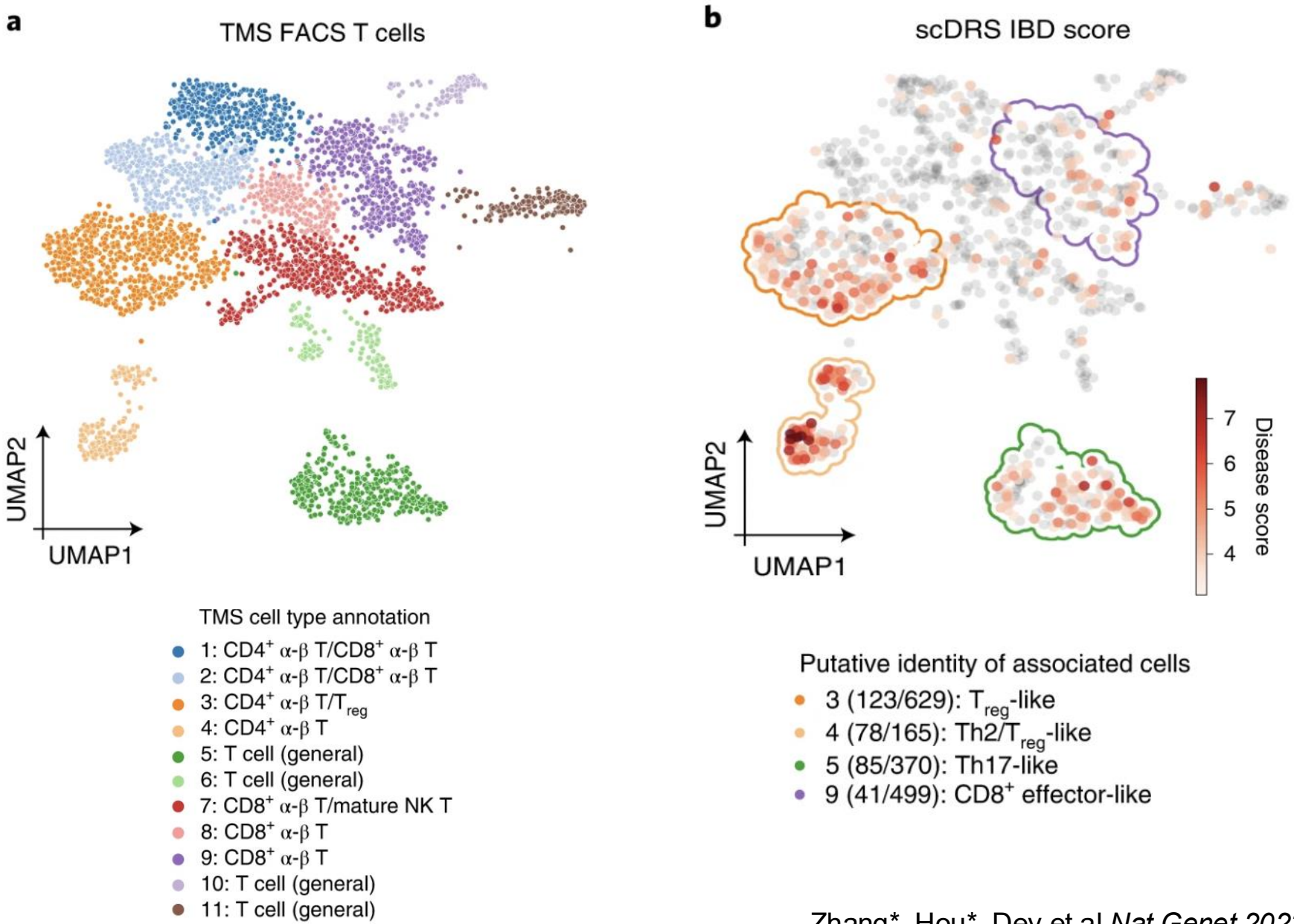


Sensitivity-specificity (SS) index: Average difference in disease signal between boxed pairs of cell types and traits versus all other pairs.

SS index

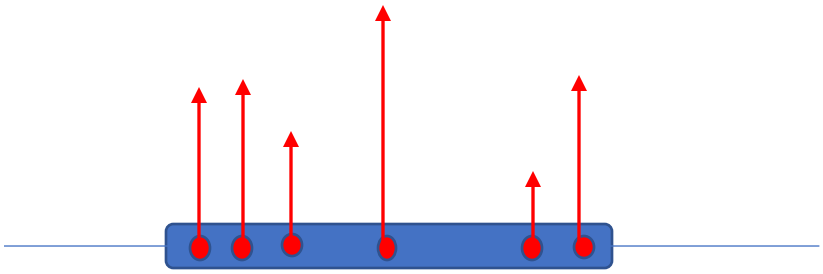


Scoring a single cell for disease association using scDRS



Zhang*, Hou*, Dey et al *Nat Genet 2022*

Prioritizing genes for a complex disease (MAGMA and PoPS)



Two types of gene test statistics have been implemented in MAGMA:

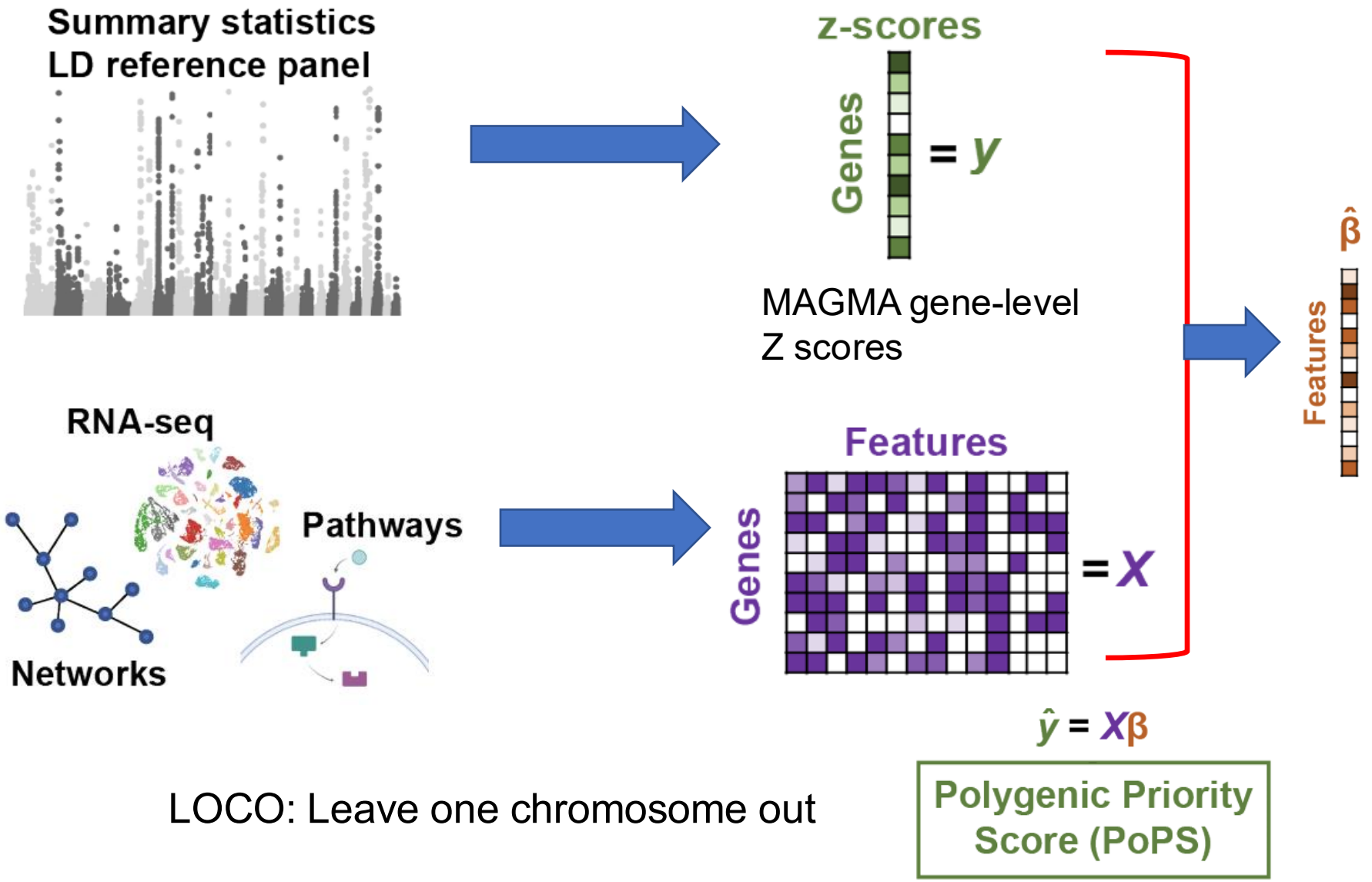
(a) The mean of the χ^2 statistic for the SNPs in a gene,

(b) The top χ^2 statistic among the SNPs in a gene.

For the mean χ^2 statistic, a gene p-value is then obtained by using a known approximation of the sampling distribution

Prioritizing genes for a complex disease (MAGMA and PoPS)

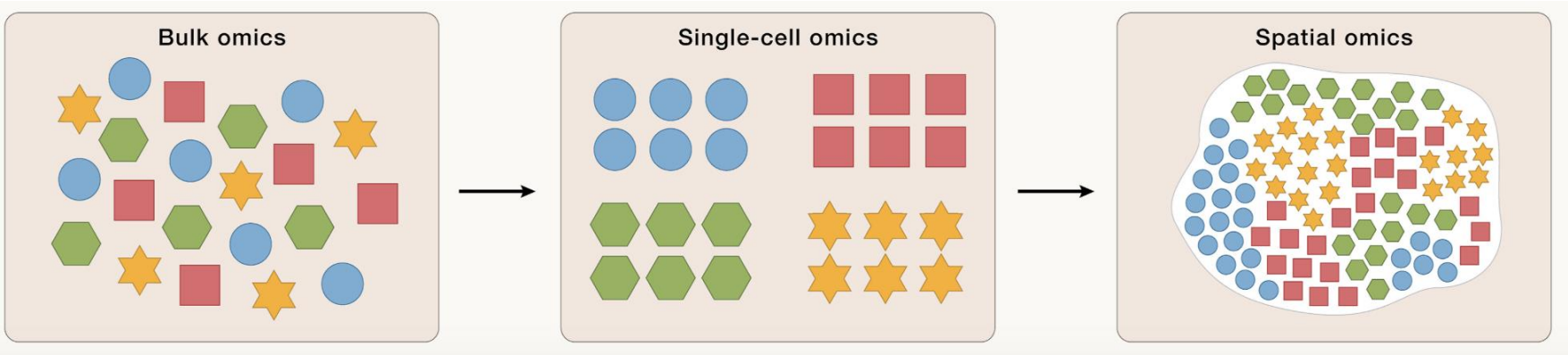
Weeks et al 2024 *Nat Genet*



LOCO: Leave one chromosome out

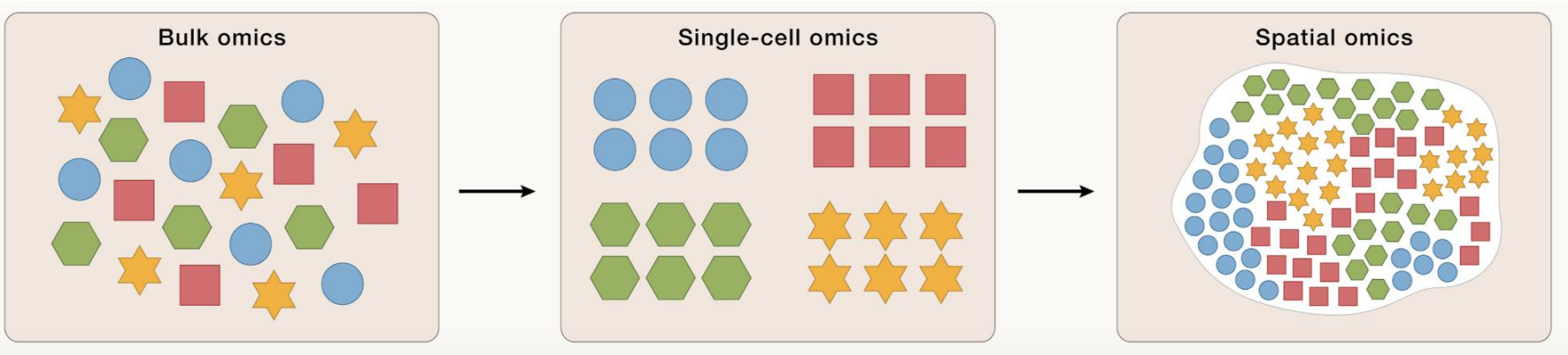
Linking spatial transcriptomics to GWAS disease risk

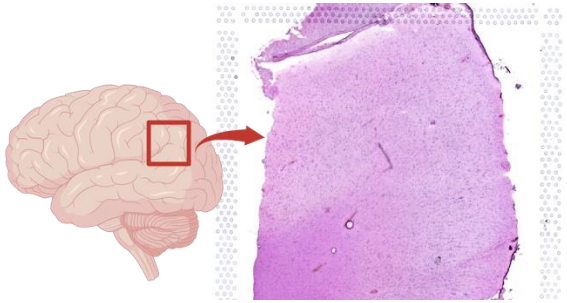
Spatial transcriptomics assays demonstrate the structure of cellular organization in a tissue

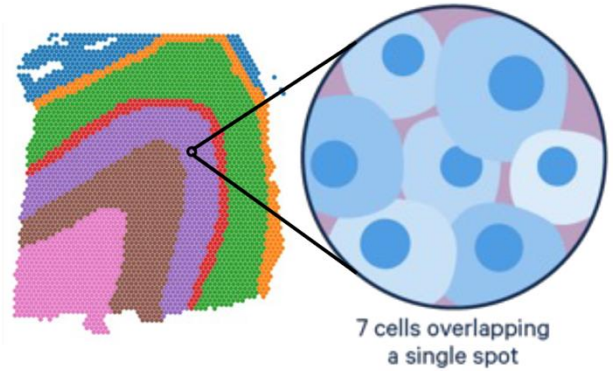


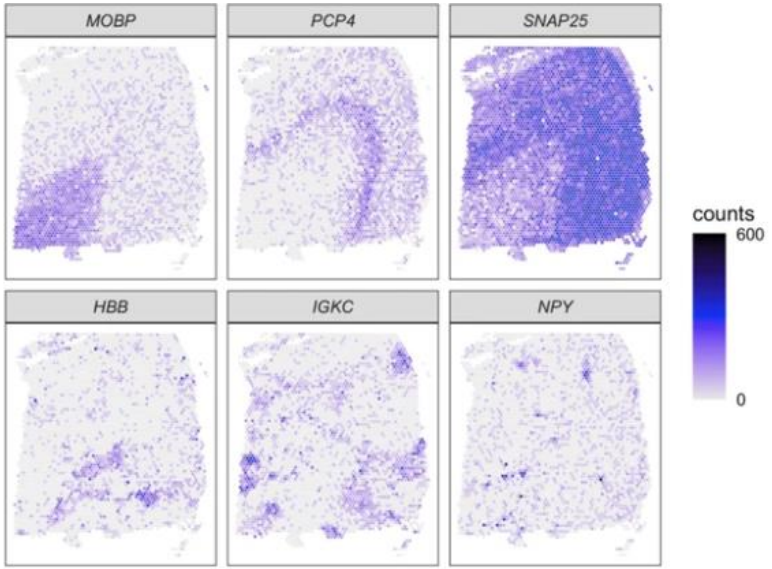
Liu et al 2024 *Cell*
Weber et al 2023 *Nat Commun*

Spatial transcriptomics assays demonstrate the structure of cellular organization in a tissue



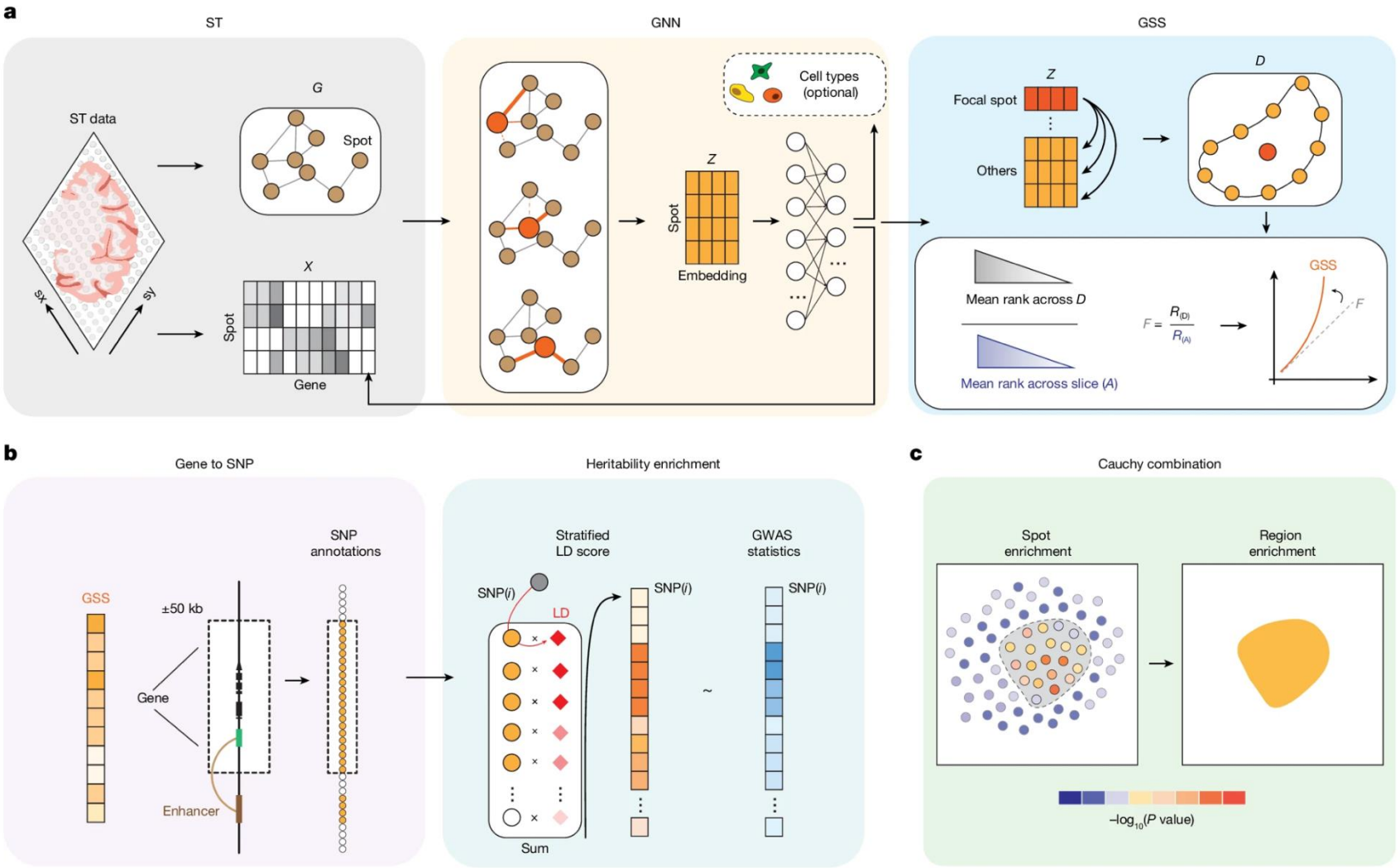






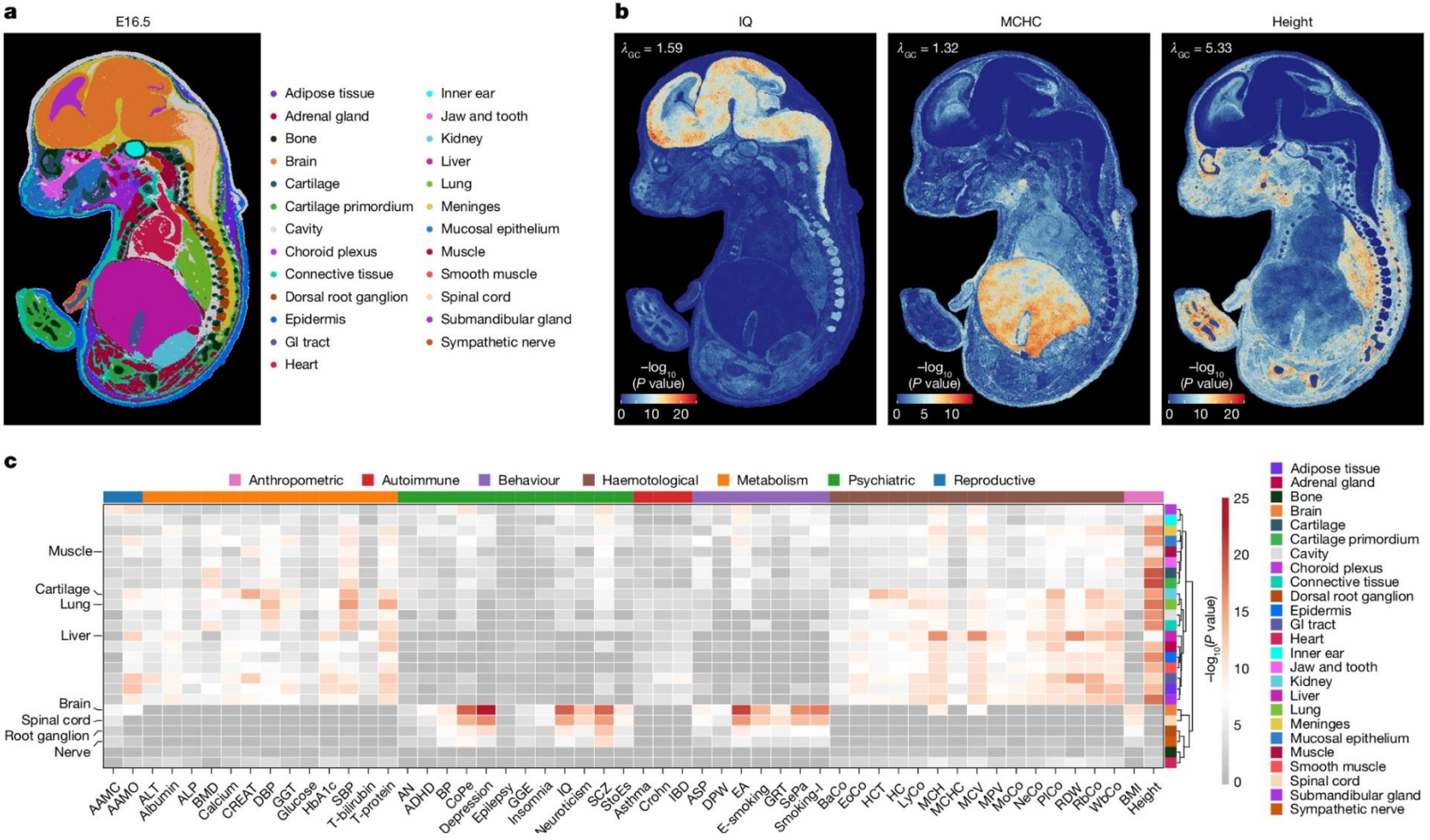
Liu et al 2024 *Cell*
Weber et al 2023 *Nat Commun*

gsMap: Scoring spatial cell-resolution disease maps



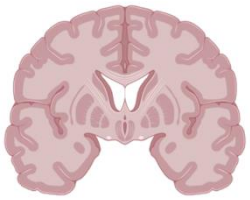
Song et al 2025 *Nature*

gsMap: Scoring spatial cell-resolution disease maps



Song et al 2025 *Nature*

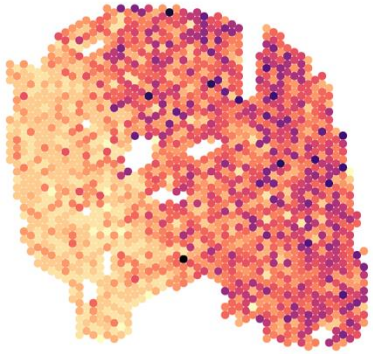
Spatial variability patterns of a gene changes between healthy and disease states of a tissue

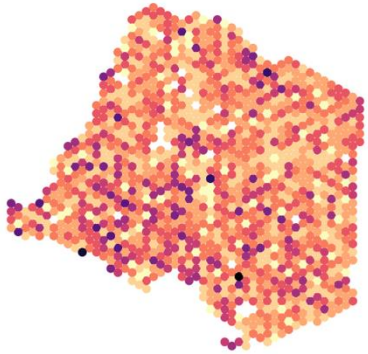


Healthy/Control sample

Alzheimer's disease sample

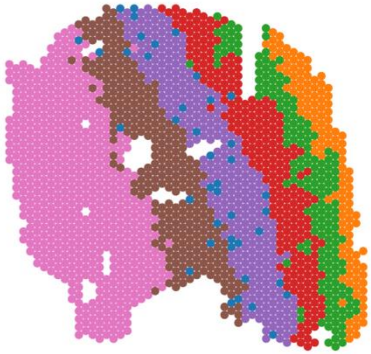
Expression of a gene

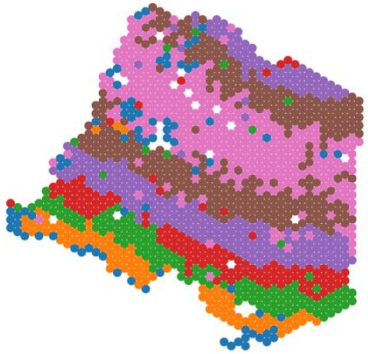




Gene expression

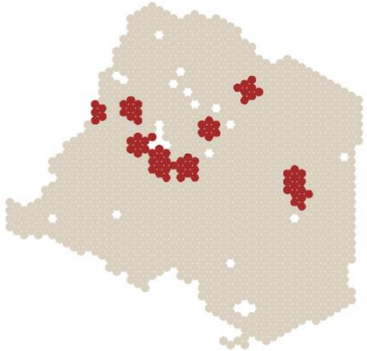
Known layer/zonal structure in brain cortex





Different layers in brain cortex

Pathology

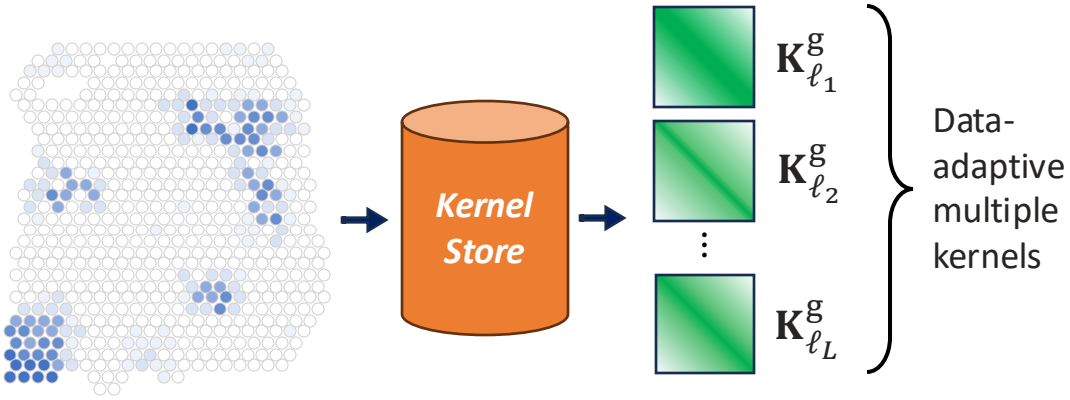


Neuritic plaques

Spacelink: Lee…Dey 2025 bioRxiv, in rev Nat Commun

Multi-scale spatial modeling of gene expression patterns can be modeled using data-driven kernels

Modeling of gene expression



$\mathbf{K}^{g}_{l}$: Exponential kernel, i.e., $\left[\mathbf{K}^{g}_{l}\right]_{a,b} = \exp\left(-\frac{\text{distance between spot a and b}}{\text{lengthscale } l}\right)$

Model: $\mathbf{y}_g = \underbrace{\boldsymbol{\mu}_g}_{\text{Fixed mean}} + \underbrace{\mathbf{U}^{g}_{1} + \cdots + \mathbf{U}^{g}_{L}}_{\text{Spatial random effects}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{Nugget effect}};$

$\mathbf{U}^{g}_{l} \sim \boldsymbol{N}(\mathbf{0}, \sigma^{g}_{l}\mathbf{K}^{g}_{l})$

$\boldsymbol{\epsilon}_g \sim \boldsymbol{N}(\mathbf{0}, \tau^{g}\mathbf{I})$

Estimate σ^{g}_{l} using the method of moments estimator of covariance.

Identify SVG by testing

H_0: $\sigma^{g}_{l} = 0$ for all l

H_a: $\sigma^{g}_{l} > 0$ for some l

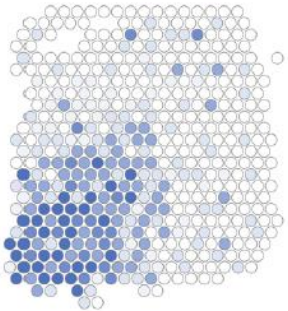
Defining a score per gene that characterizes the effective spatial variability of a gene in a tissue

$$\mathbf{y}_g = \underbrace{\boldsymbol{\mu}_g}_{\text{Fixed mean}} + \underbrace{\mathbf{U}_1^g + \cdots + \mathbf{U}_L^g}_{\text{Spatial random effects}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{Nugget effect}};$$

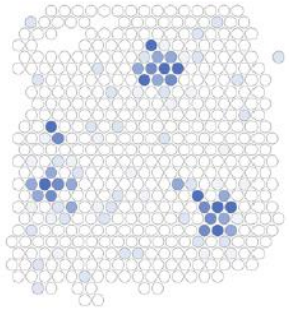
$$\mathbf{U}_l^g \sim \boldsymbol{N}(\mathbf{0}, \sigma_l^g \mathbf{K}_l^g)$$
$$\boldsymbol{\epsilon}_g \sim \boldsymbol{N}(\mathbf{0}, \tau^g \mathbf{I})$$

$$\rightarrow \text{ESV} = \frac{\sum_{l=1}^{L} w_l \sigma_l^g}{\sum_{l=1}^{L} \sigma_l^g + \tau^g}$$

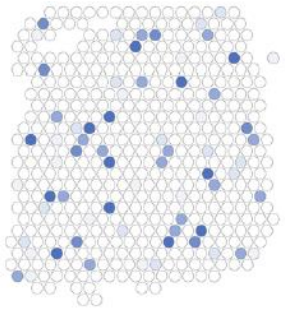
$$w_l = \frac{\left\| \mathbf{K}_l^g - \mathbf{I} \right\|_F}{\left\| \mathbf{K}_l^g \right\|_F} \in [0,1]$$



Global SVG with high ESV

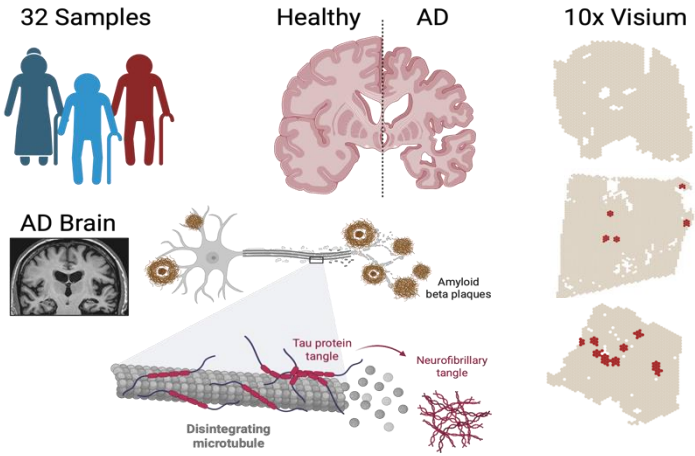


Global SVG with low ESV



Not global SVG

Evaluating the performance of ESV on Alzheimer's spatial data along different pathology stages

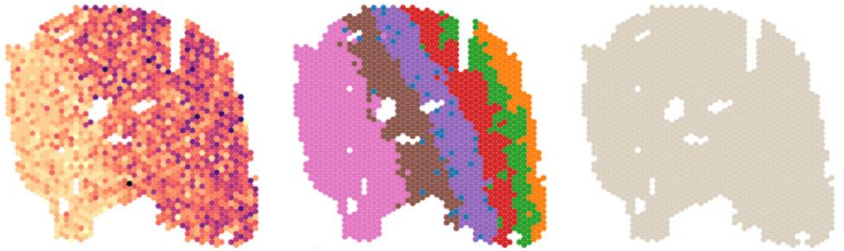


Pathological measures:

- Total amyloid plaque burden
- Neurofibrillary tangles accumulation
- Number of neuritic plaques

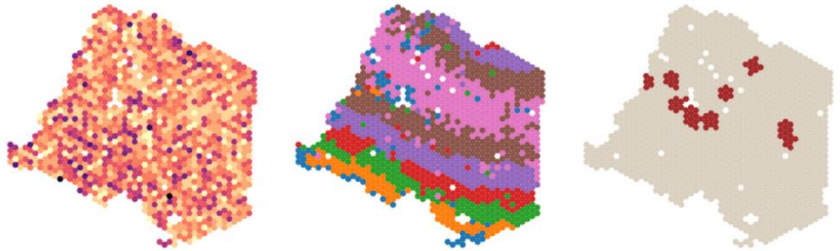
PKM

Control:

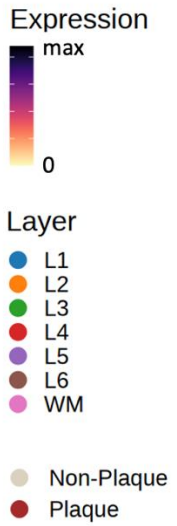


Amyloid (sqrt): 0.2727, ESV: 0.8416

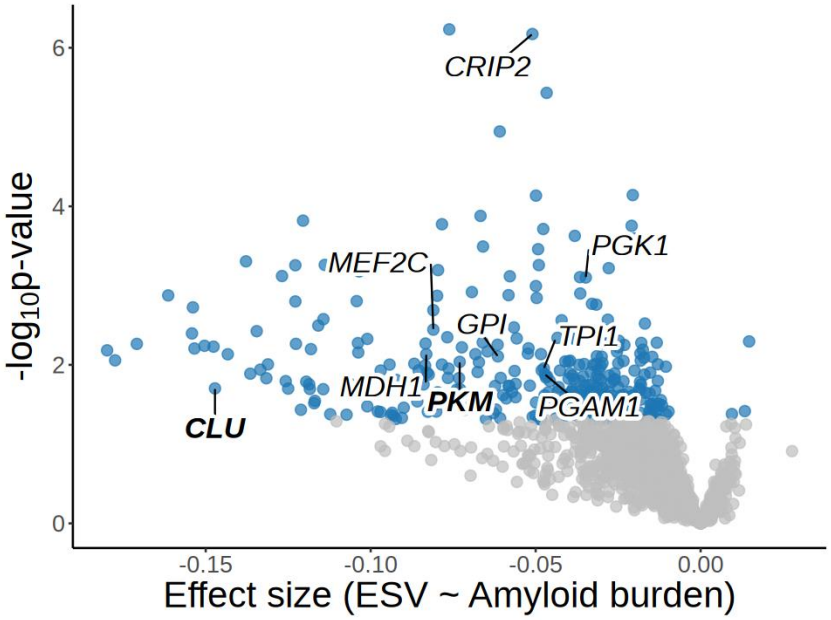
Later AD stage:

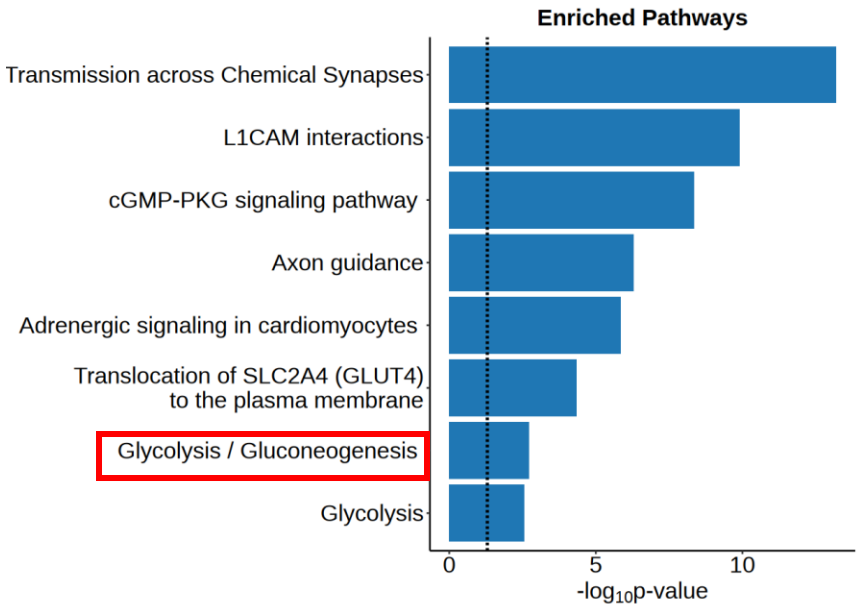


Amyloid (sqrt): 3.7748, ESV: 0.0594



We see higher numbers of genes showing trends of ESV change against pathology compared to other metrics



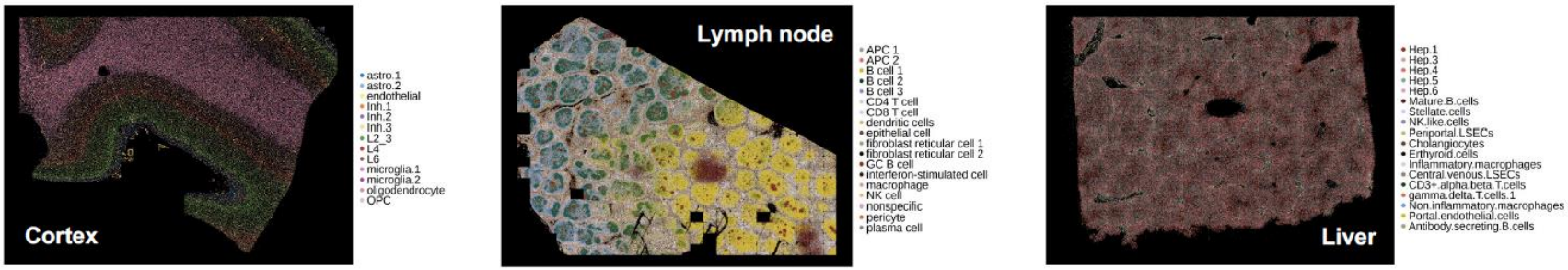


334 genes showed nominally significant ESV trends against AD disease pathology,

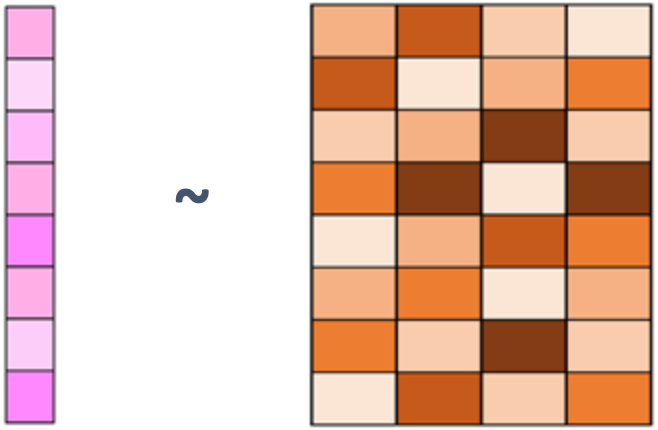
1.2-3.1x higher than other spatial gene prioritization scores.

Assessing predictive power for tissue-associated disease-related genes using Spacelink ESV features

a



MAGMA gene prioritization score based on GWAS disease association (de Leeuw et al 2015 PLoS Comp Biol)

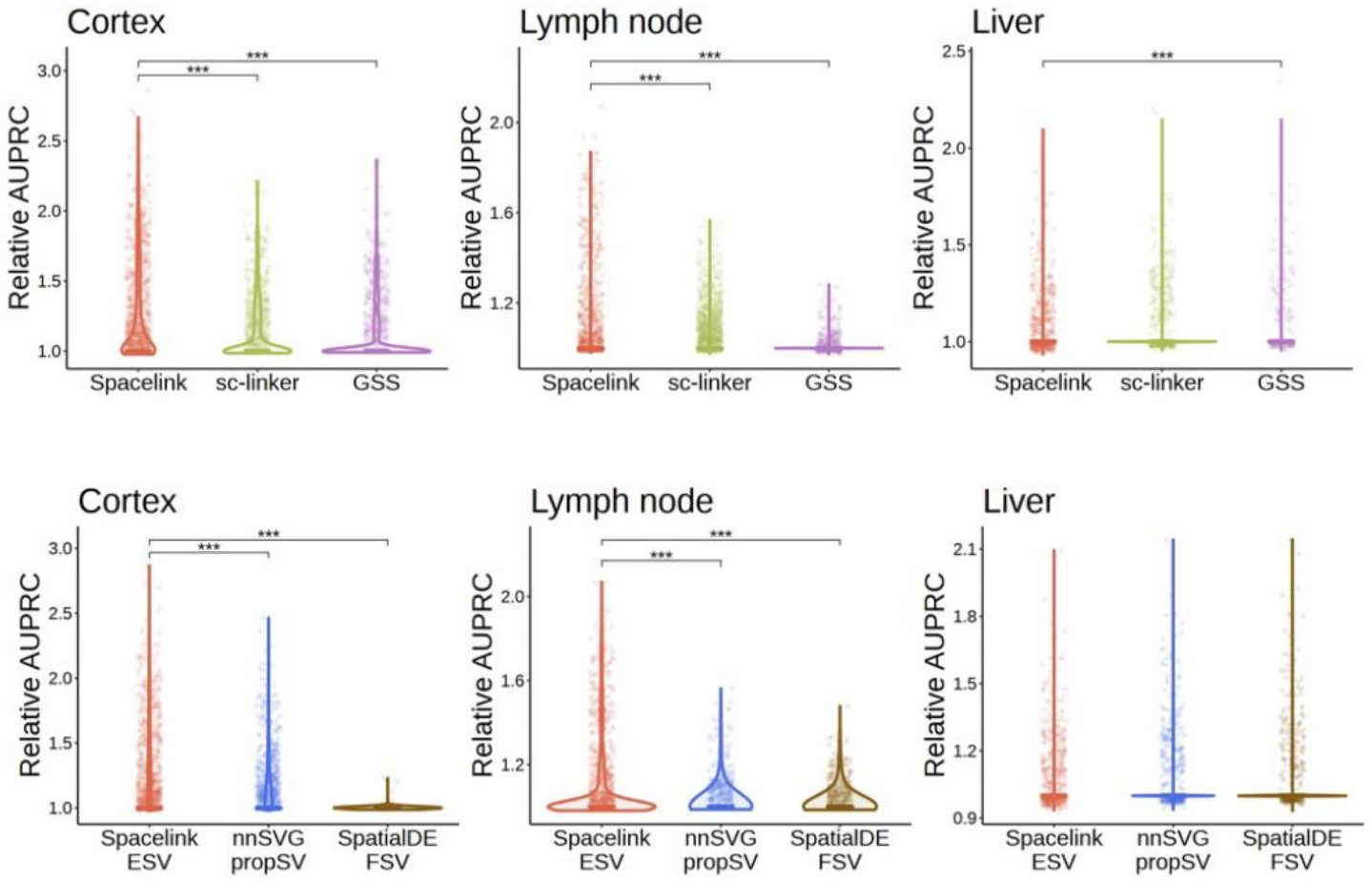


$Y \sim X$

Spacelink ESV gene scores for each cell type in a tissue + pooled across cell types

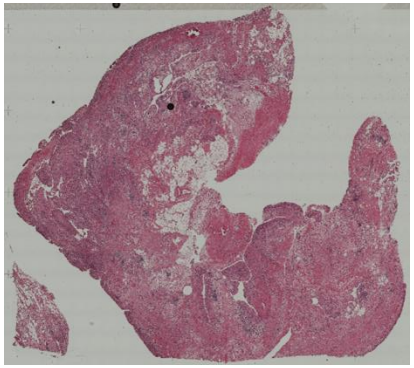
We perform this regression by leave-one-chromosome-out and predicting on the held-out chromosome to get a Spacelink integrated gene score fine-tuned for a disease.

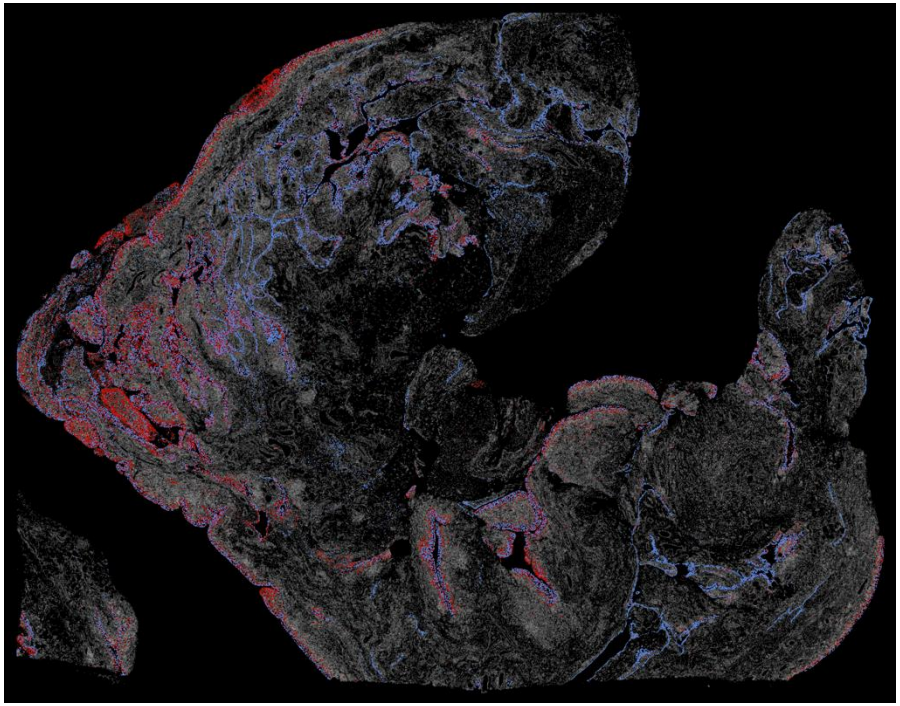
Spacelink ESV scores across different cell types outperform other spatial and non-spatial scores in disease gene prediction task



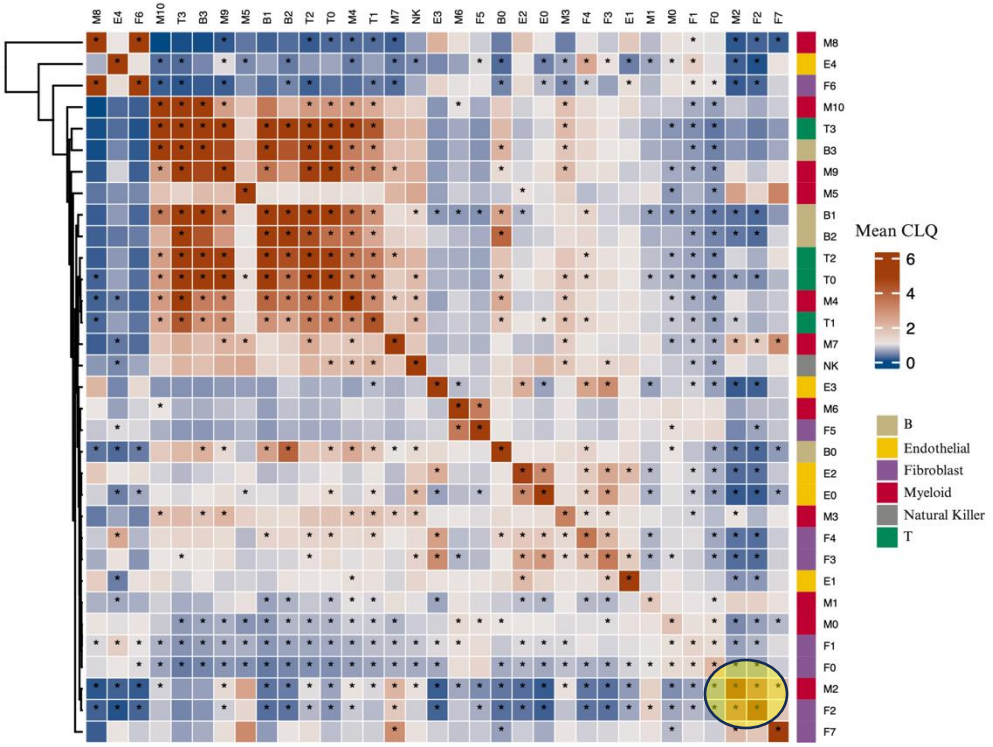
GSS: Song et al 2025 *Nature*
sc-linker: Jagadeesh*, Dey* et al 2022 *Nat Genet*

Spatial co-localization of specific cell types is observed under certain disease pathological states





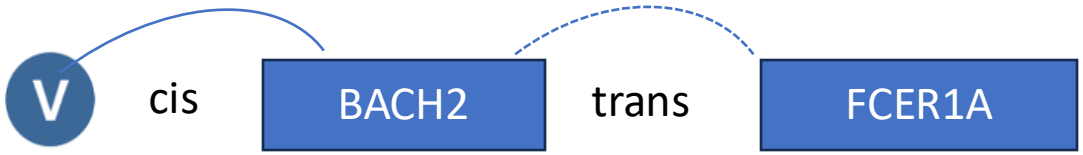
Example case of spatial colocalization of two cell types in disease pathology: SPP1-macrophages and lining fibroblasts



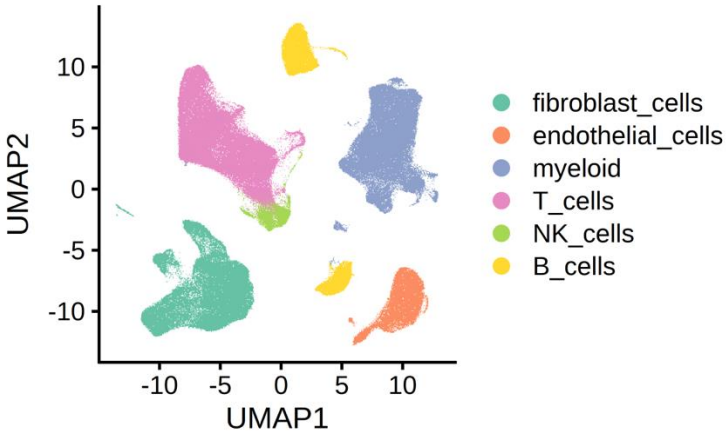
Mantel*,Zhang* et al 2025, in rev *Sci. Trans Med*

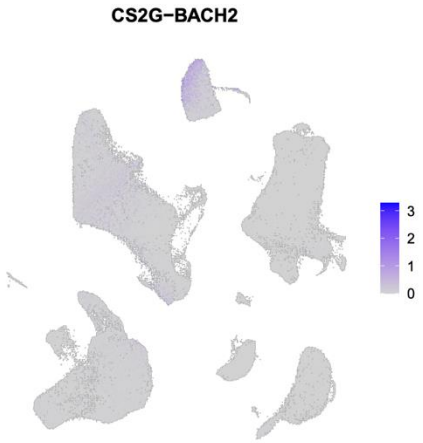
Spatial patterns in gene expression can elucidate cross cell-type cis-trans mechanisms downstream of genetic variation

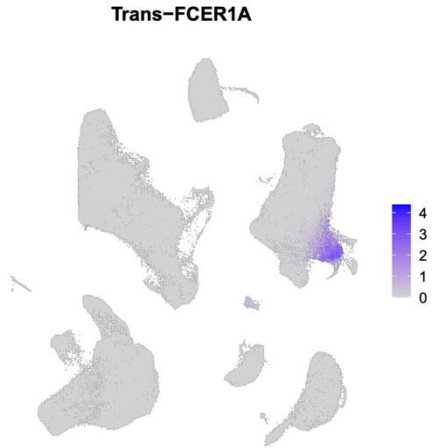
RA finemapped variant *rs72928038*

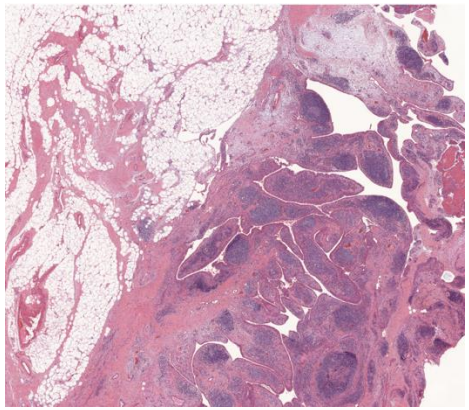


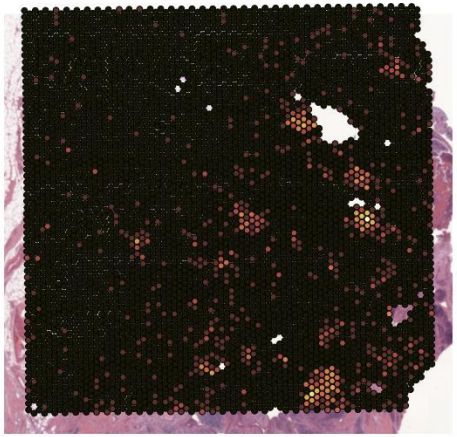
eQTLGen

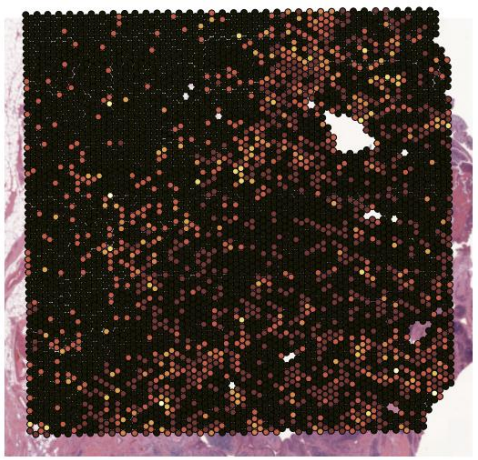












Harry Zhang (C. Leslie, L. Donlin)

Tip of the iceberg to the full iceberg