# Linking natural and artificial genomic perturbations to human disease risk

Kushal K. Dey

Assistant Member
Computational and Systems Biology
Memorial Sloan Kettering Cancer Center

# The Human Genome Project

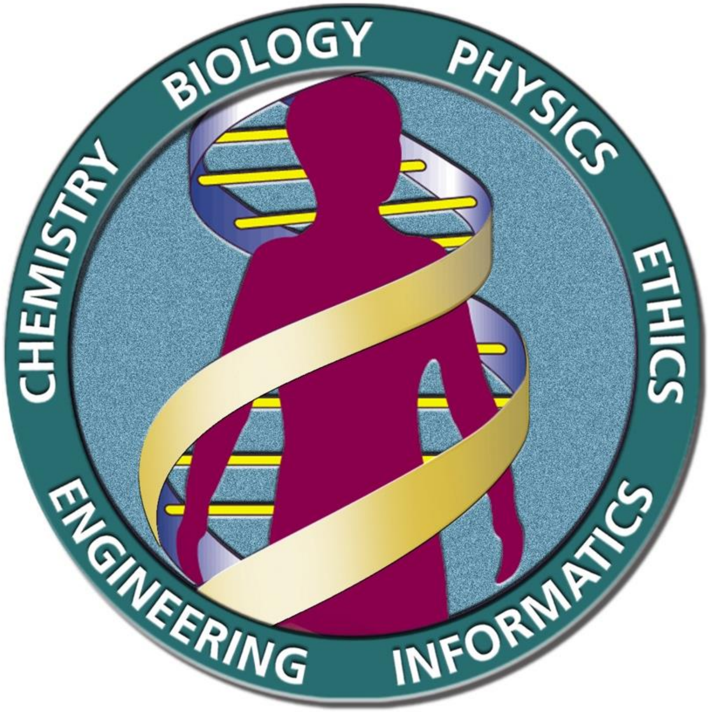
- Human Genome Project (launched 1990, completed 2003)

- 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



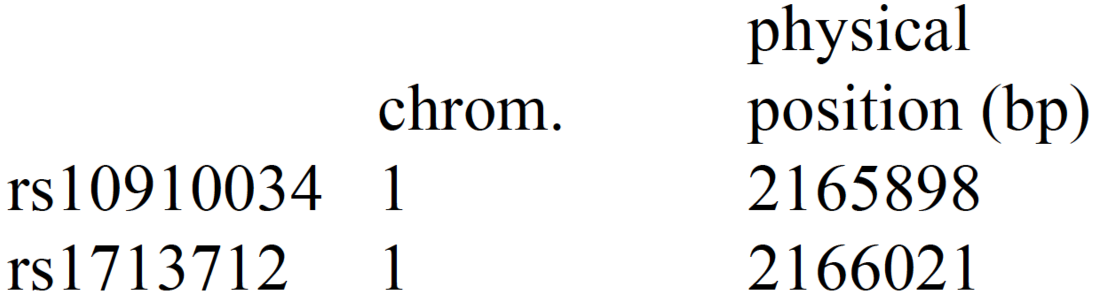
# HapMap Project: Cataloguing variations in the sequences of human DNA (2002-2010) (1,000 individuals)

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)

biobank uk
Improving the health of future generations

BioBank Japan
バイオバンク・ジャパン

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

MVP
10 YEARS
MILLION VETERAN PROGRAM

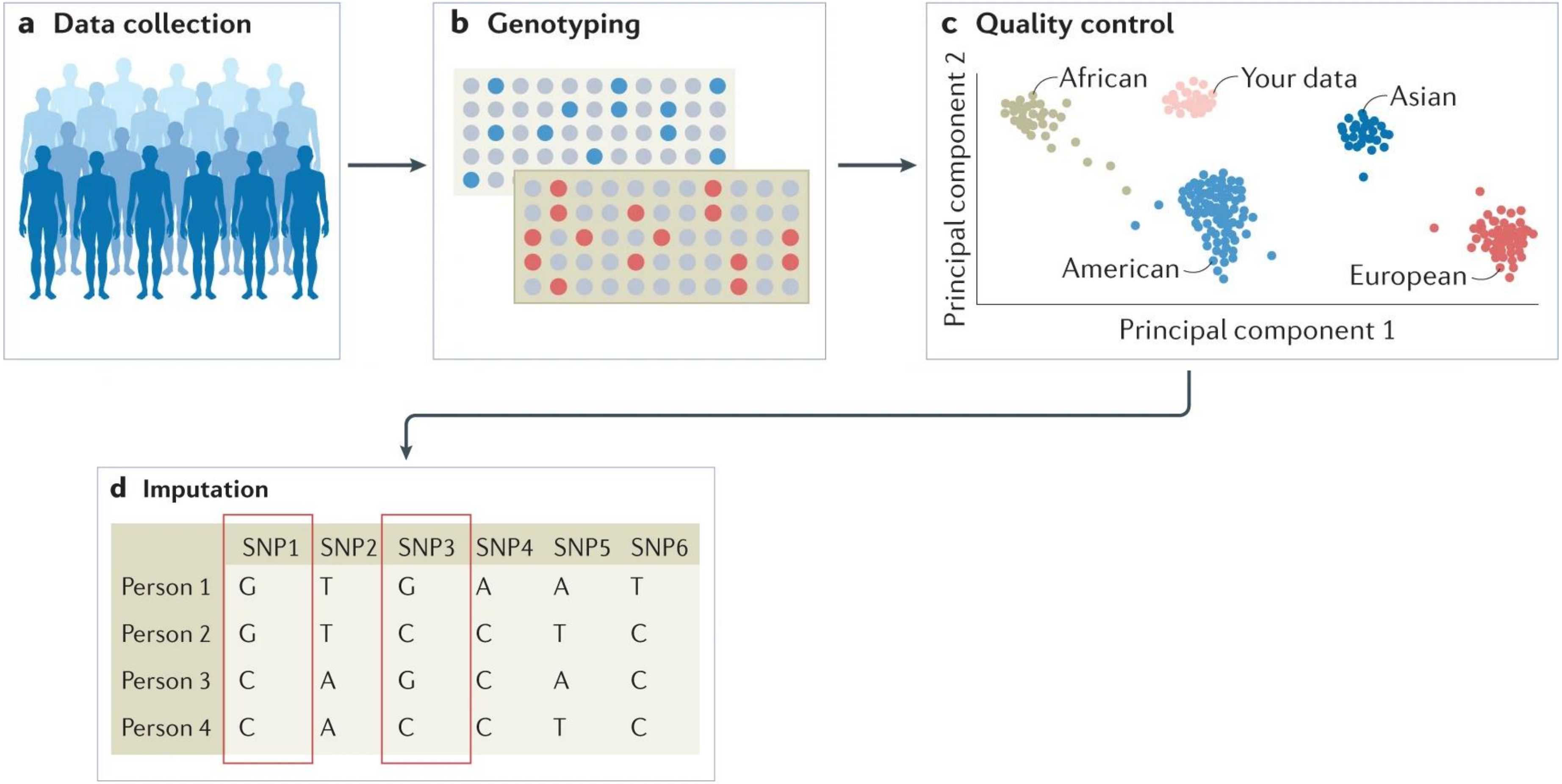
Disease-related prospective studies (~10K-100K)

PGC
Psychiatric Genomics Consortium

The COVID-19 Host Genetics Initiative

CARDIoGRAMplusC4D

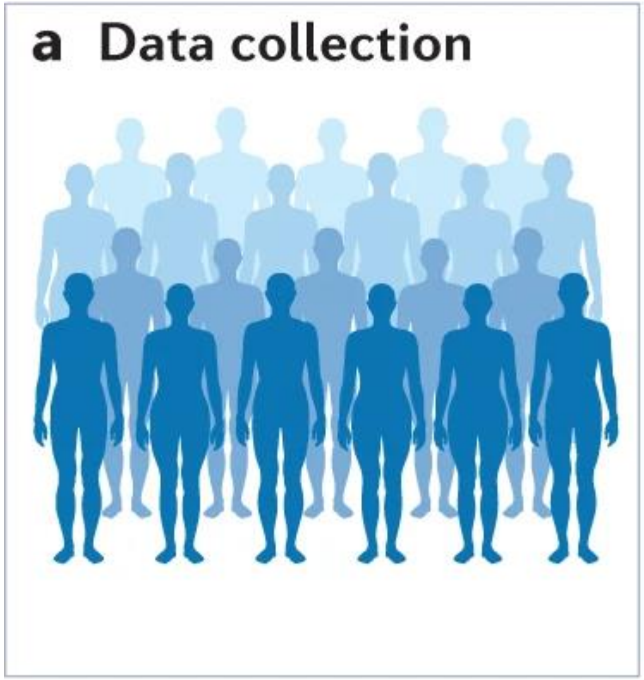
# Collecting genotype data from many many individuals ( order of 100, 000 individuals)

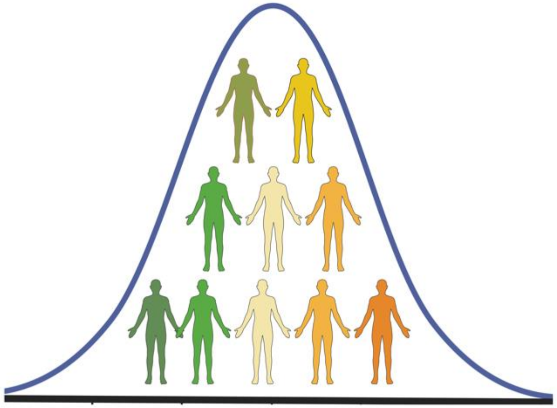


| | SNP1 | SNP2 | SNP3 | SNP4 | SNP5 | SNP6 |
|---|---|---|---|---|---|---|
| Person 1 | G | T | G | A | A | T |
| Person 2 | G | T | C | C | T | C |
| Person 3 | C | A | G | C | A | C |
| Person 4 | C | A | C | C | T | C |

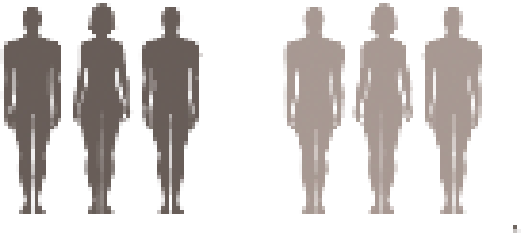
Cano-Gomez & Trynka 2020 *Front Genet*

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





Quantitative phenotype (red blood cell count, LDL cholesterol)

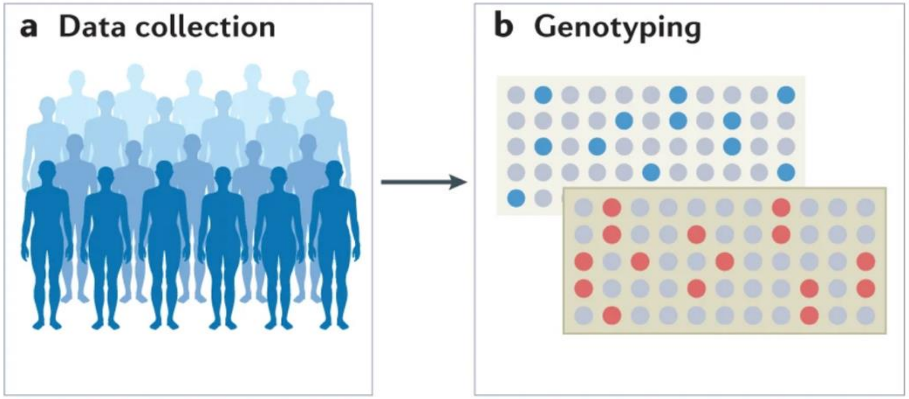


Cases

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

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

(linear or logistic regression)

Genotype effect

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

Local population relatedness

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

**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

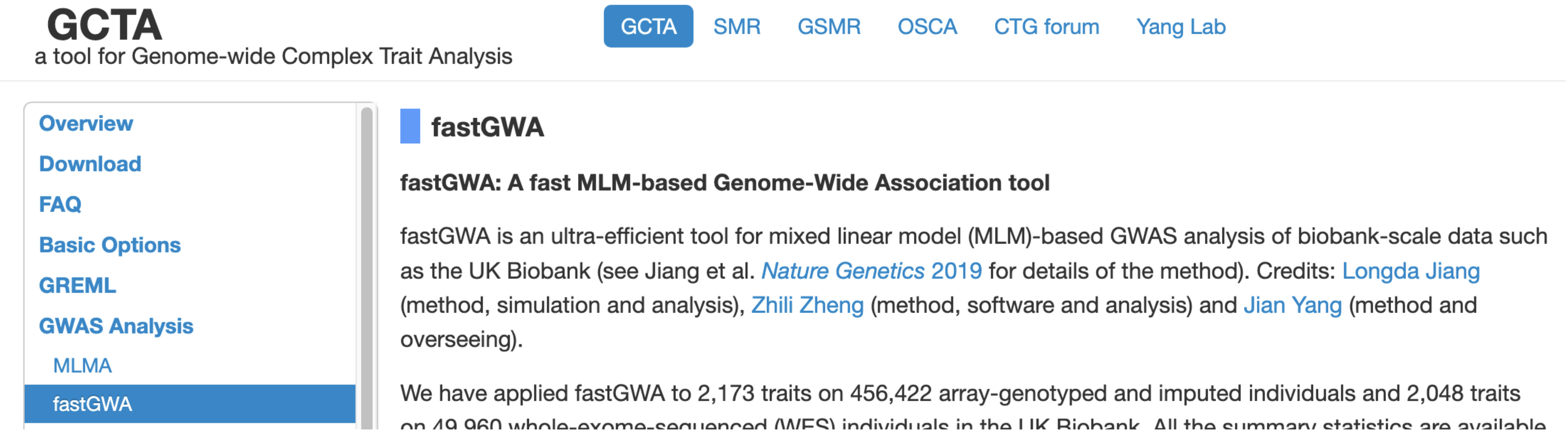
Estimates of the effect size

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

$$\boxed{\mathrm{V} = \sigma_g^2\,\psi + \sigma_e^2 I}$$

Overall phenotypic variance-covariance matrix = genetic + error

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



## 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

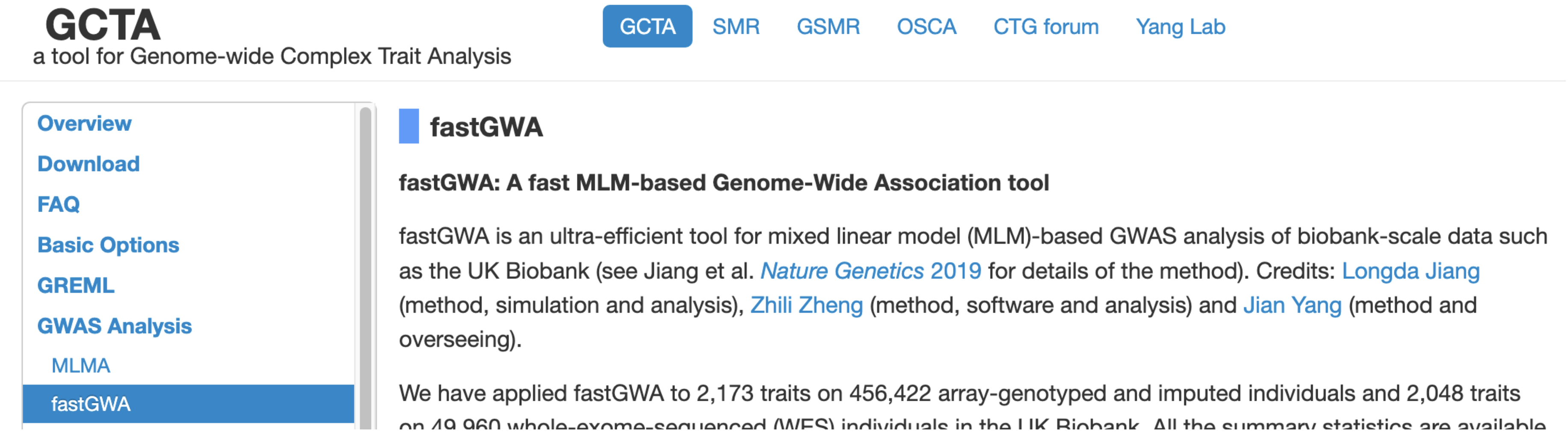
Estimates of the effect size

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

$$\boxed{\mathrm{V} = \sigma_g^2 \, \psi + \sigma_e^2 I}$$

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).

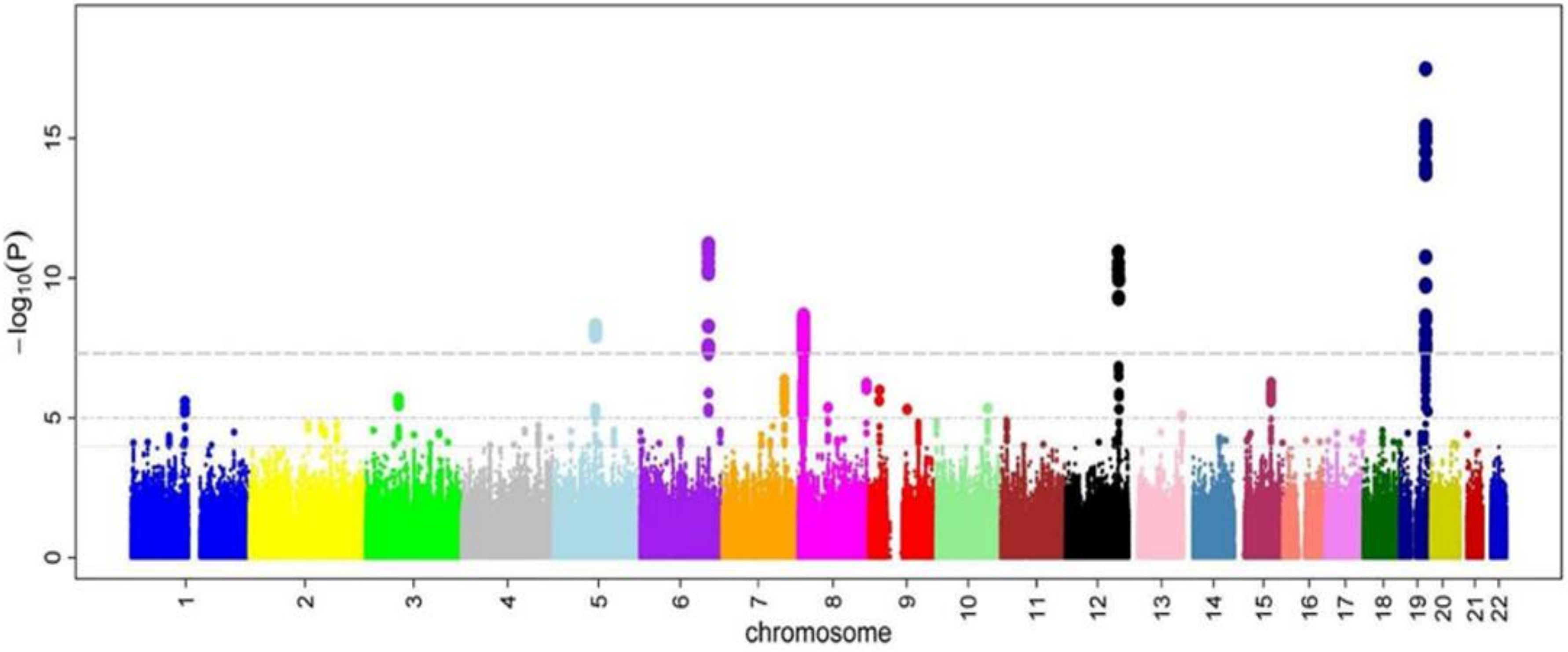
We have applied fastGWA to 2,173 traits on 456,422 array-genotyped and imputed individuals and 2,048 traits on 49,960 whole-exome-sequenced (WES) individuals in the UK Biobank. All the summary statistics are available

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

Jiang et al 2019 *Nat Genet*

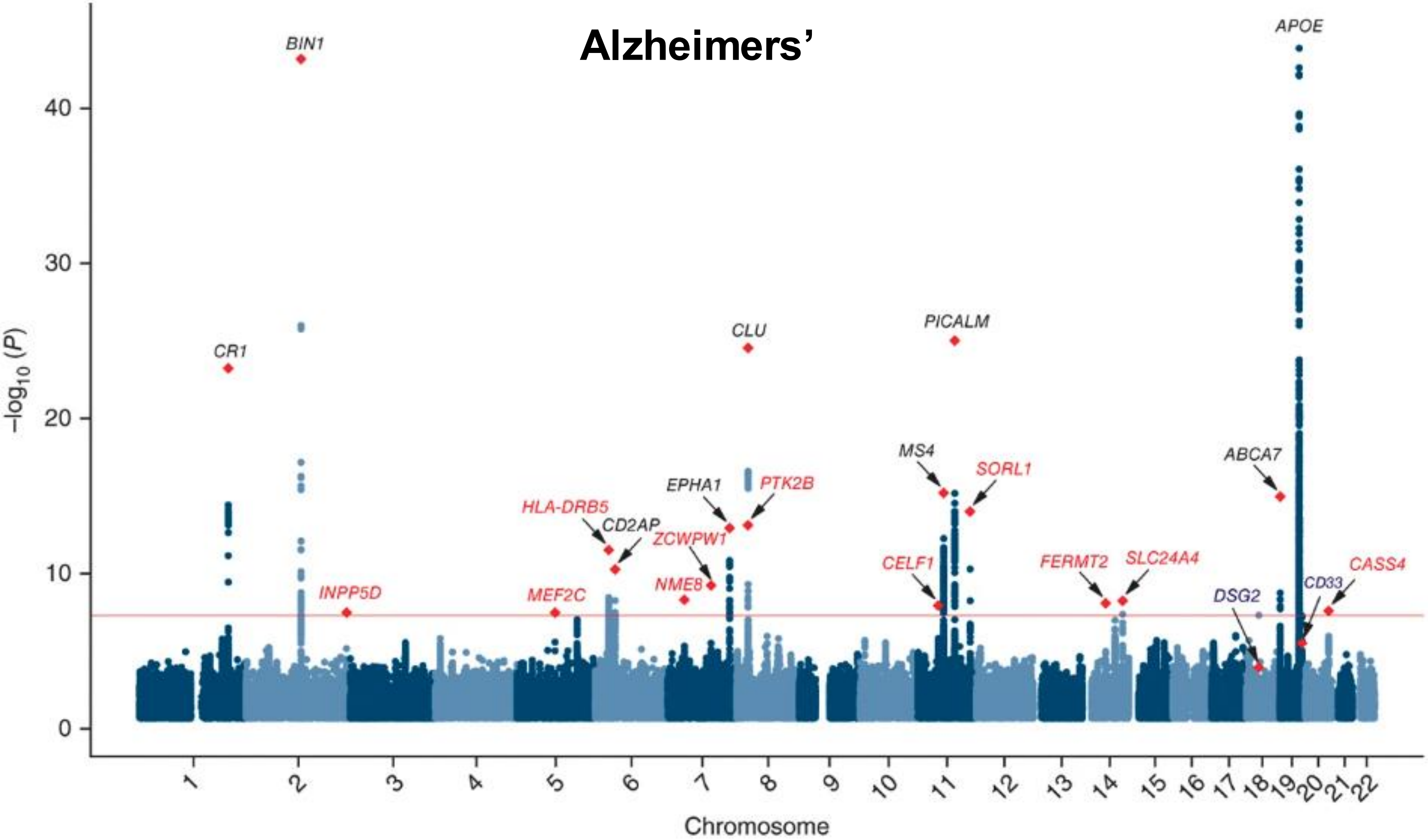
# Standard visualization technique for GWAS results

**Schizophrenia**



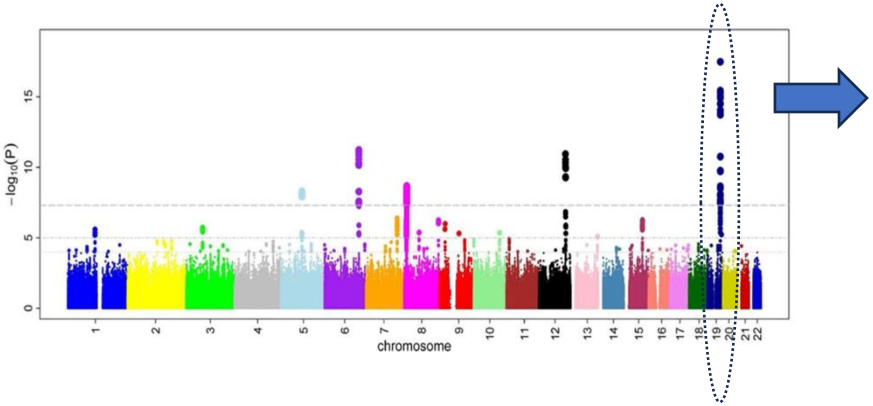
SCZ WG of PGC Consortium, 2014, *Nature*

# Standard visualization technique for GWAS results

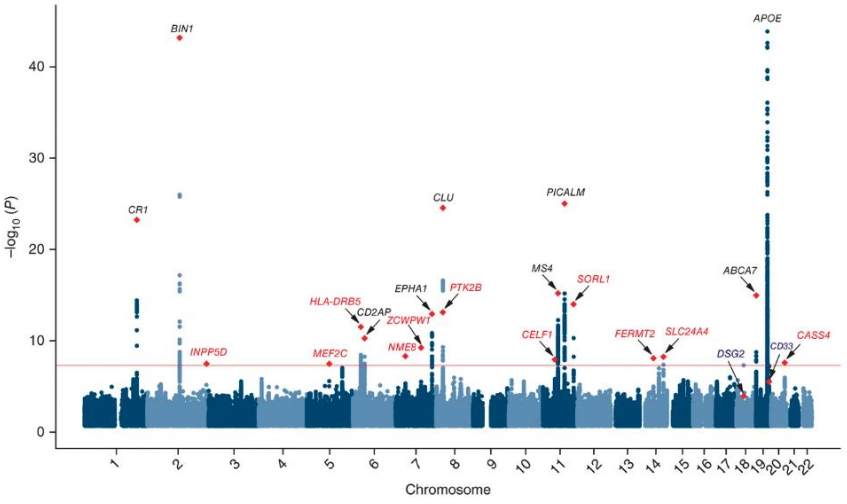


Lambert et al 2013, *Nat Genet*

# 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.



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

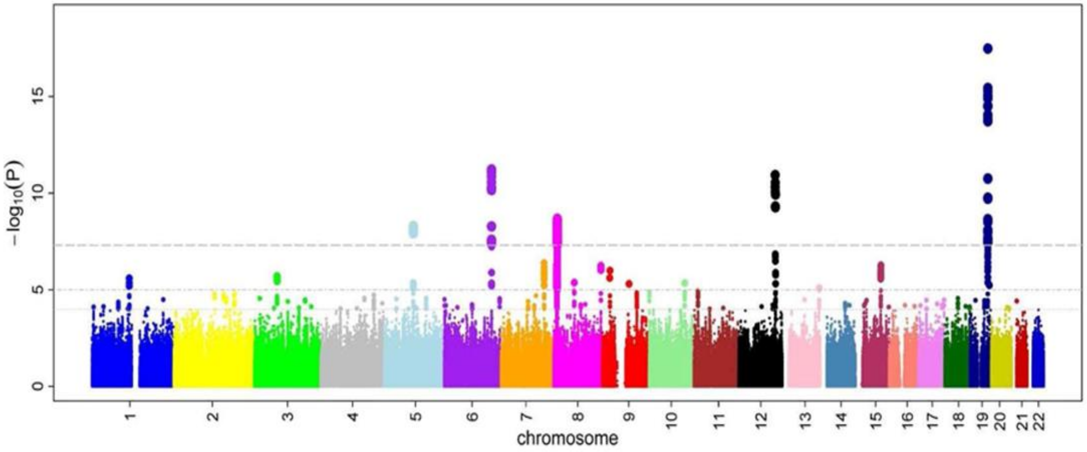
**Heritability of GWAS hits ($\boldsymbol{h^2_{GWAS}}$):** Squared correlation between best fit linear model of all GWAS hits and the phenotype

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

**Heritability of all SNPs ($\boldsymbol{h^2_g}$):** Squared correlation between best fit linear model of all SNPs and the phenotype

$$\max_w [\ r^2 (\textstyle\sum_s w_s X_{ns}, Y_n)\ \ ]$$

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



## Schizophrenia

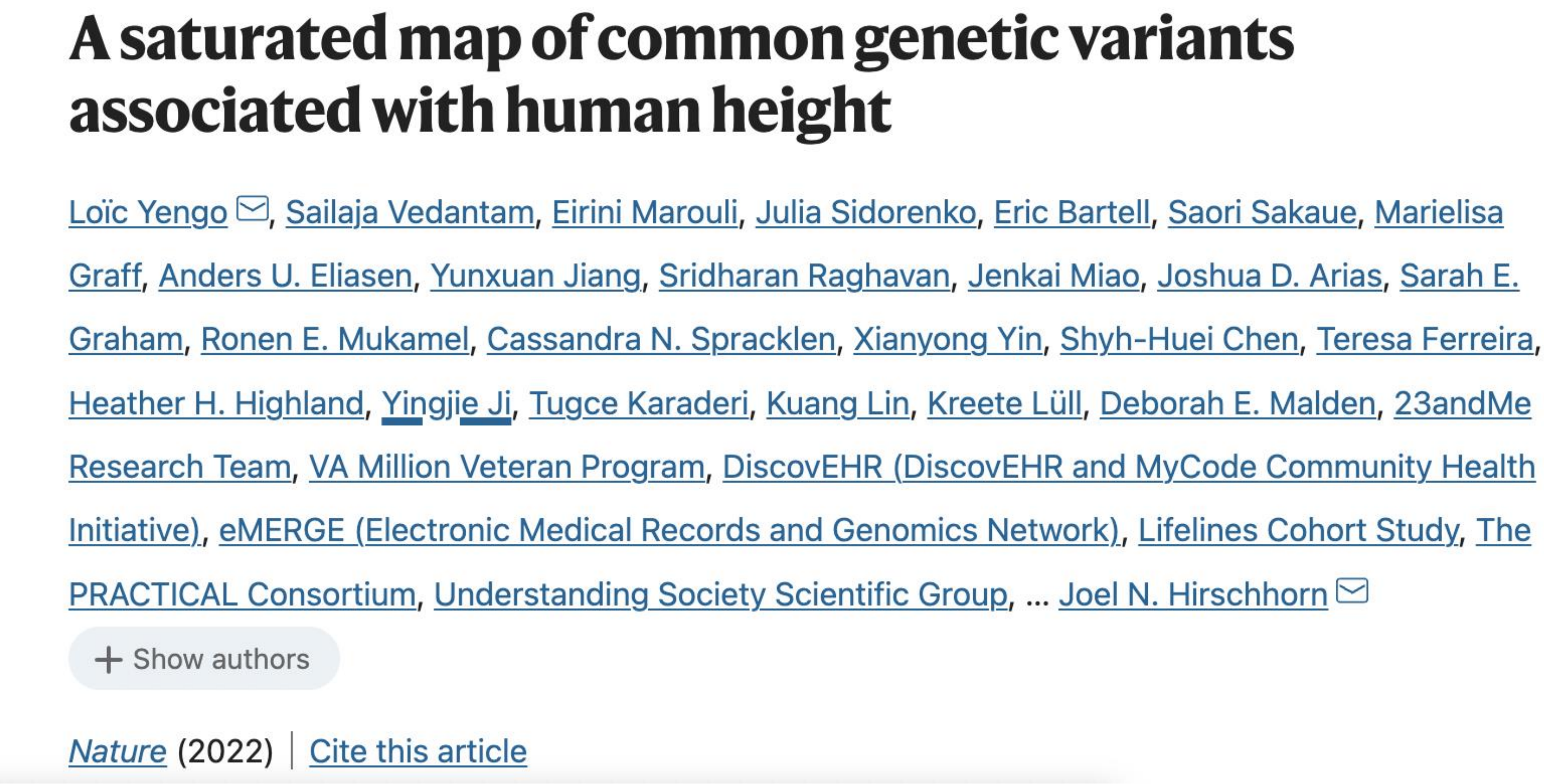
0.07 < 0.24

$h^2_{GWAS}$ $h^2_g$

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*

# The magnitude of GWAS significance depends on the LD structure around variants

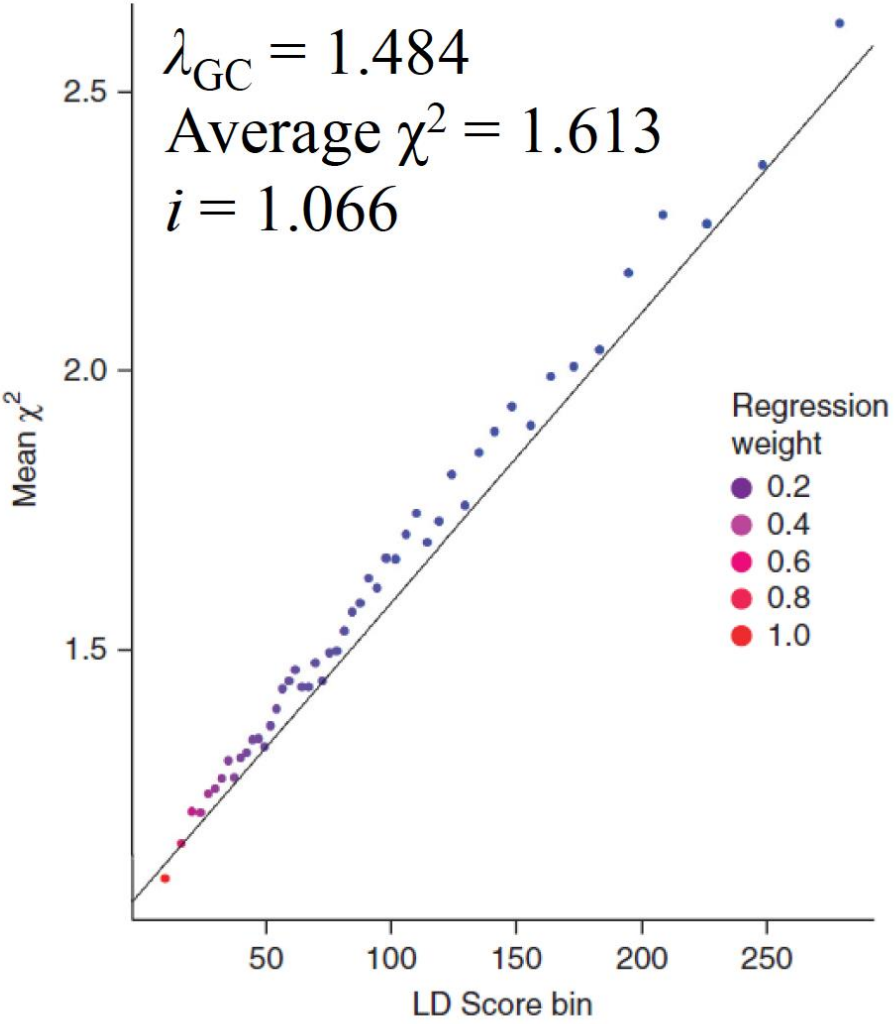
Intuition: SNPs in higher LD with other SNPs tend to have larger test statistics on average for a polygenic trait, because of more causal variants being tagged.

LDscore (SNP *x* )
= $\sum_m r^2\,(x, m)$

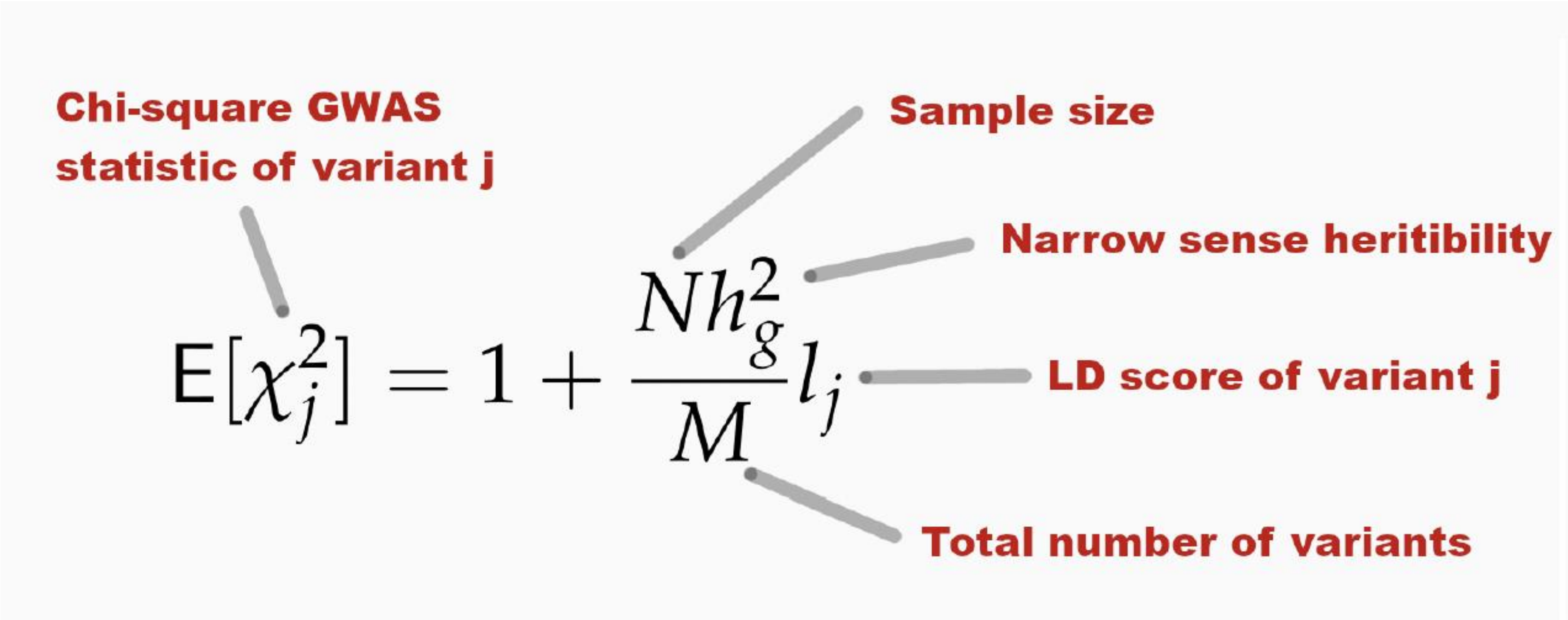
$\chi^2$ = squared Z score

*r* is the correlation between the genotypes $X_{nm}$ and $X_{nx}$

## LD score regression



# Mathematical overview of LD score regression



$$\mathrm{E}[\chi_j^2] = 1 + \frac{Nh_g^2}{M} l_j$$

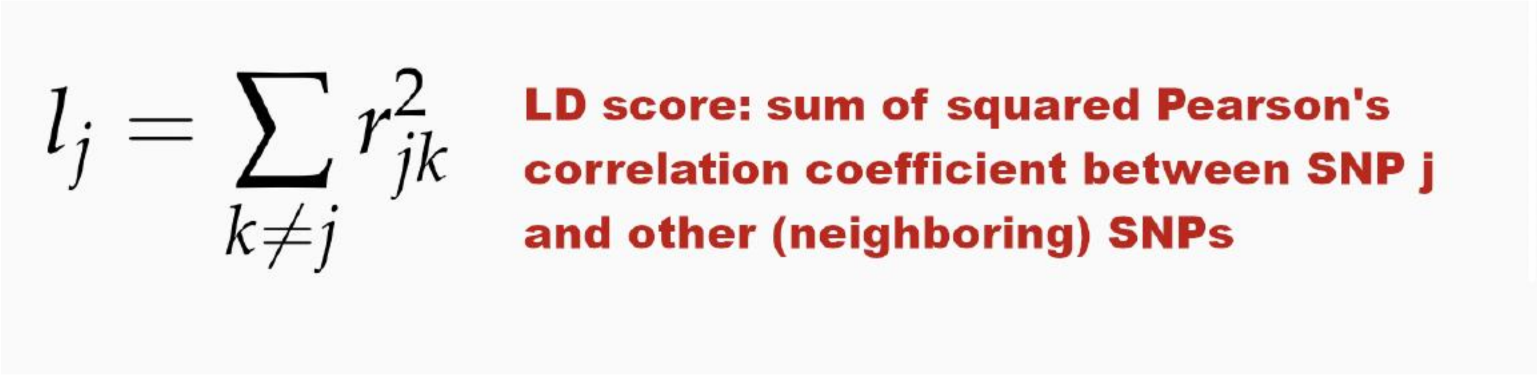
Chi-square GWAS statistic of variant j

Sample size

Narrow sense heritibility

LD score of variant j

Total number of variants



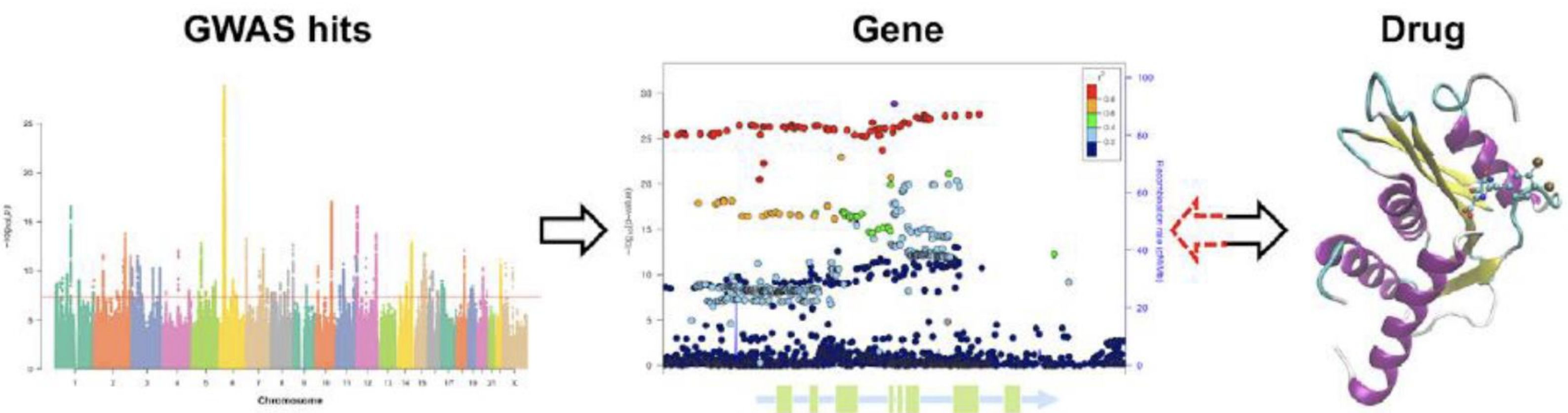
$$l_j = \sum_{k \neq j} r_{jk}^2$$

LD score: sum of squared Pearson's correlation coefficient between SNP j and other (neighboring) SNPs

Bulik-Sullivan Loh et al, 2015, *Nat Genet*

# Clinical and therapeutic implications of GWAS

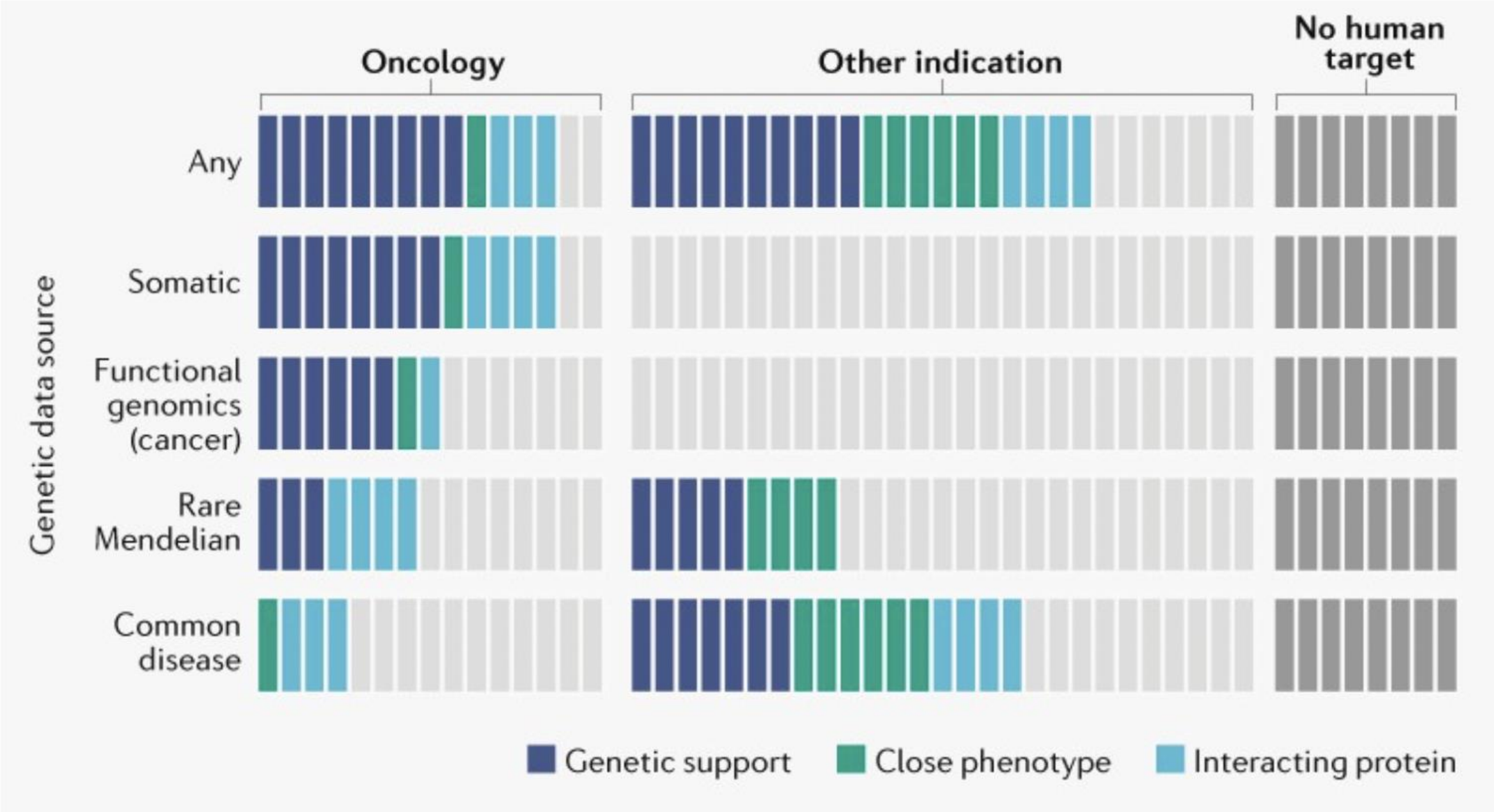
# Is GWAS actually important? (GWAS hits to drugs)



| 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)

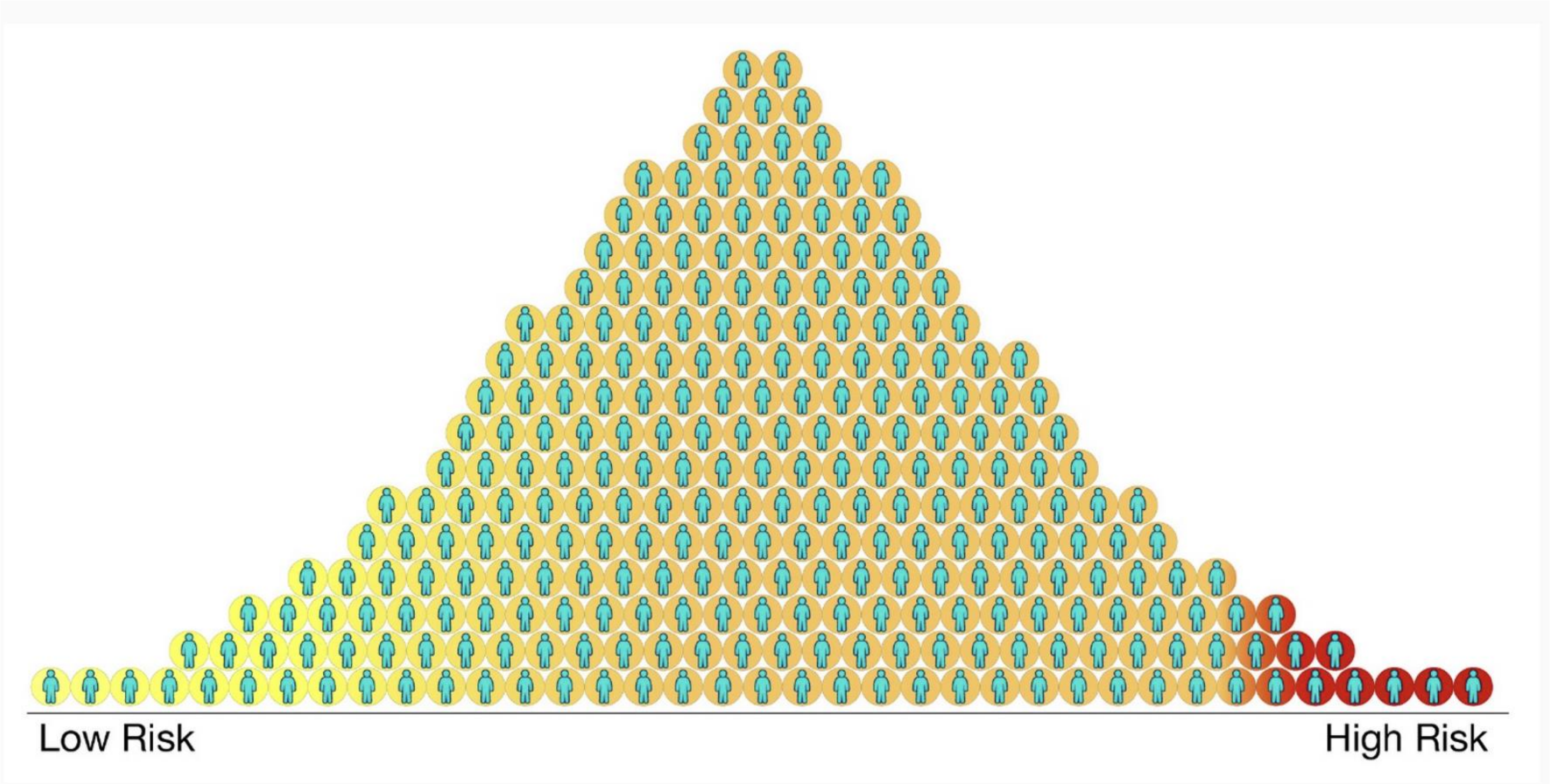


33 of 50 FDA approved drugs in 2021 have genetic support, with highest implicated from common disease GWAS.

Ochoa et al 2022 *Nat Rev Drug Disc.*

# Is GWAS actually important? (Genetic risk score)

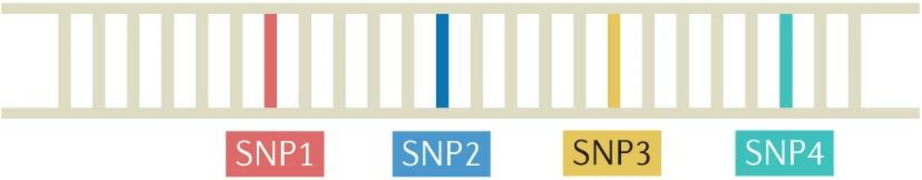
Identify the genetic risk for any individual for diseases and traits based on their genetic make-up (genotypes across all SNPs). Are they at risk for a specific disease?



# How to calculate polygenic risk scores?

① **GWAS summary statistics**

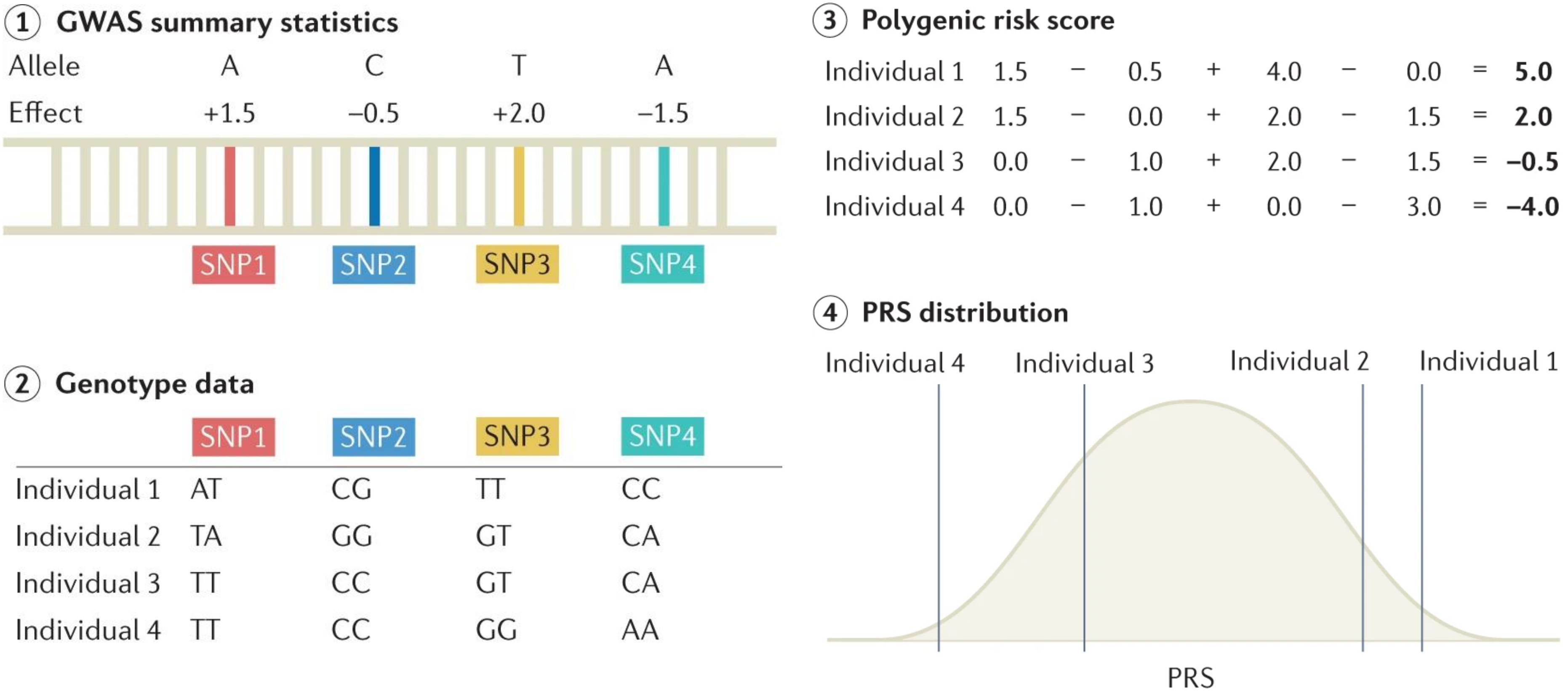
| Allele | A | C | T | A |
|---|---|---|---|---|
| Effect | +1.5 | −0.5 | +2.0 | −1.5 |



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 |

# 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**

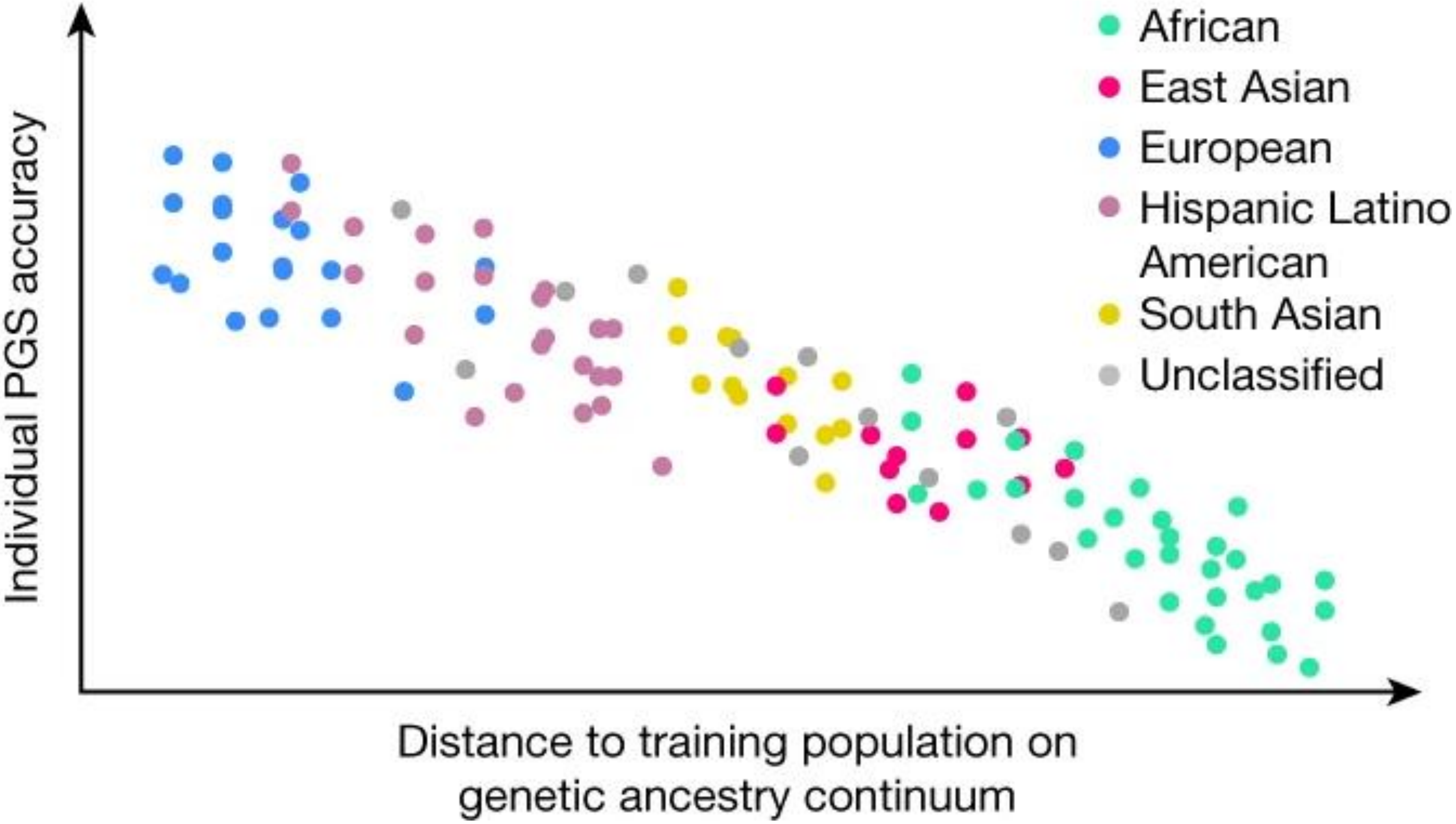
| | SNP1 | SNP2 | SNP3 | SNP4 |
|---|---|---|---|---|
| Individual 1 | AT | CG | TT | CC |
| Individual 2 | TA | GG | GT | CA |
| Individual 3 | TT | CC | GT | CA |
| Individual 4 | TT | CC | GG | AA |

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**

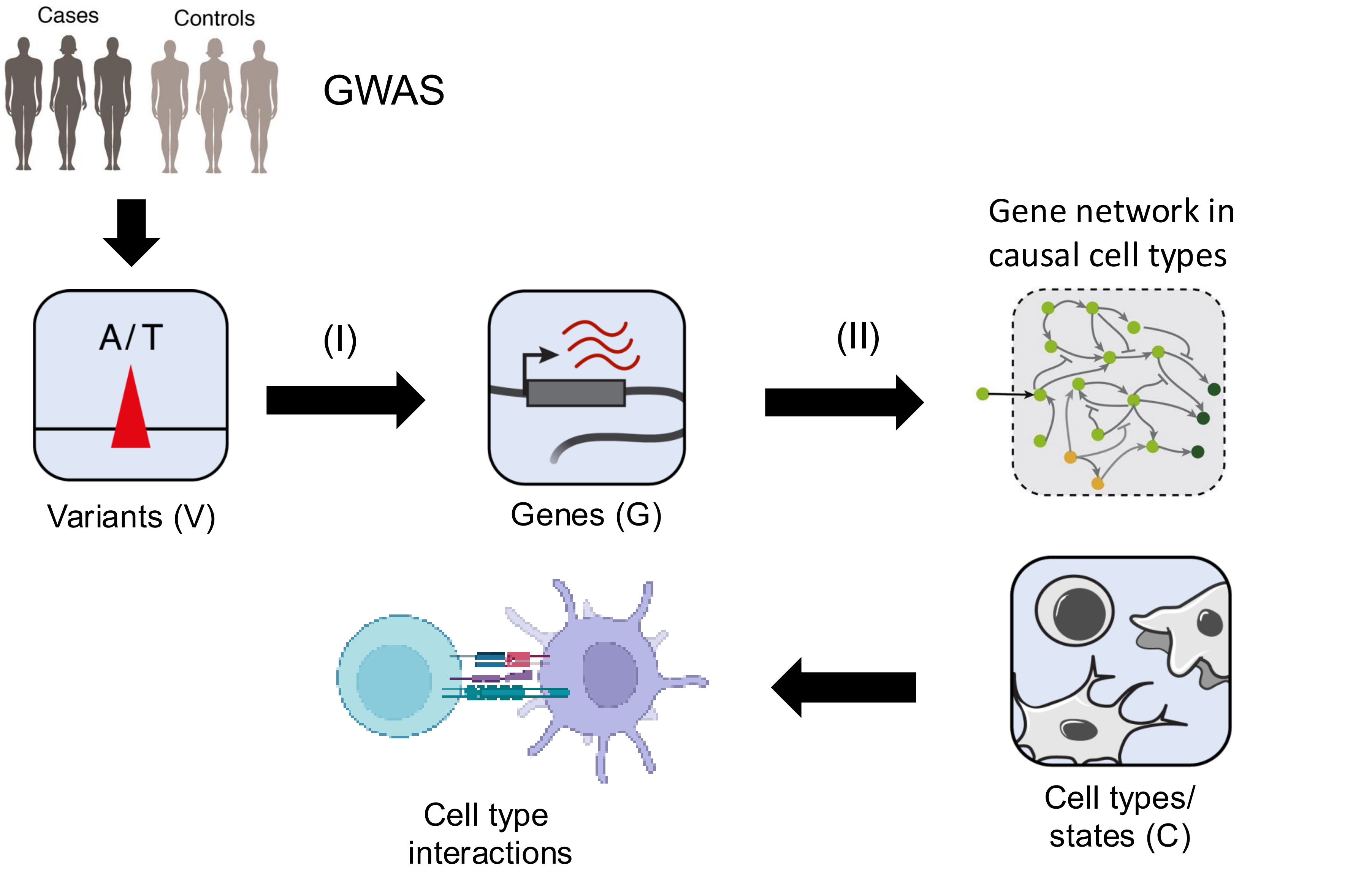
# A big challenge in polygenic risk scores (representation)



Ding et al 2023 *Nature*

# GWAS-to-function (Overview)

# Understanding the functional basis of GWAS variants



GWAS

(I)

(II)

Gene network in causal cell types

Variants (V)

Genes (G)

Cell type interactions

Cell types/ states (C)

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

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

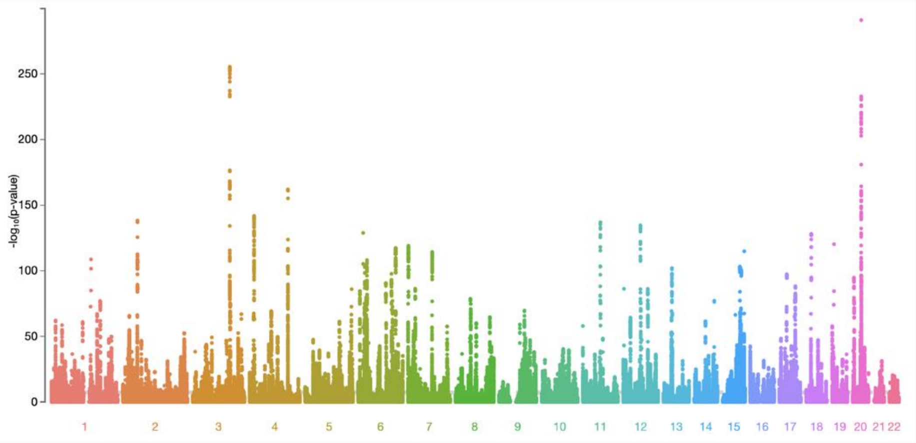
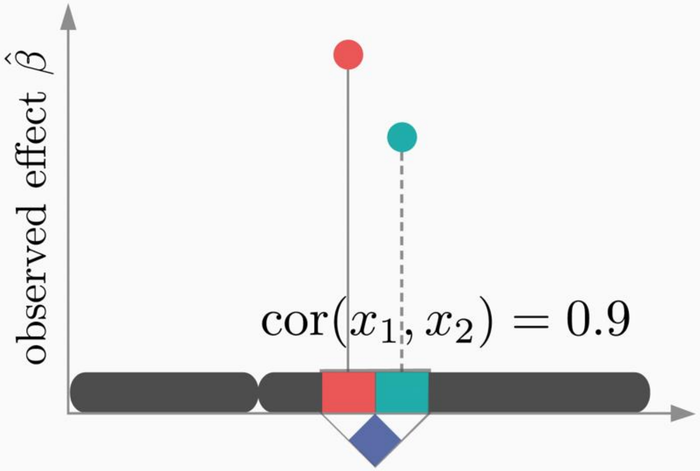


**Figure:** UK Biobank height GWAS, `http://nealelab.is/uk-biobank`



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

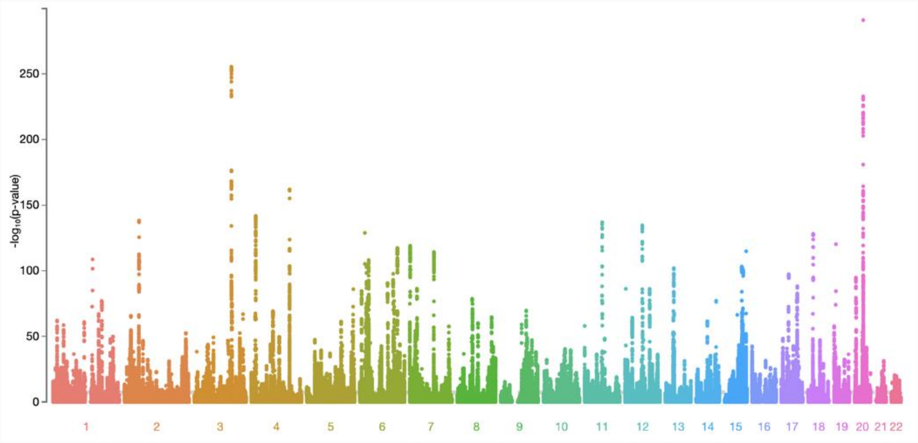
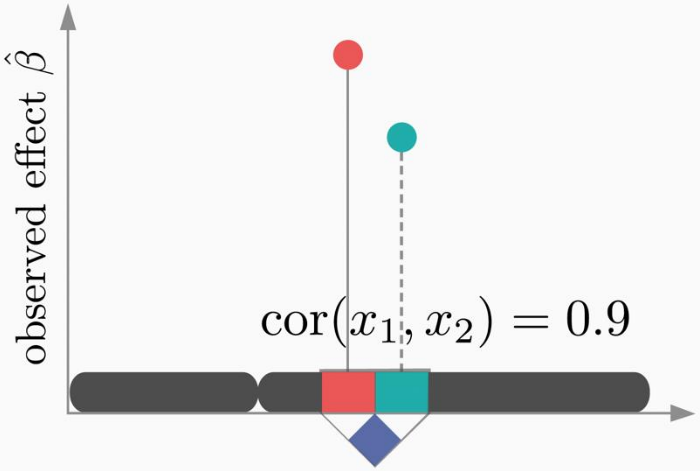
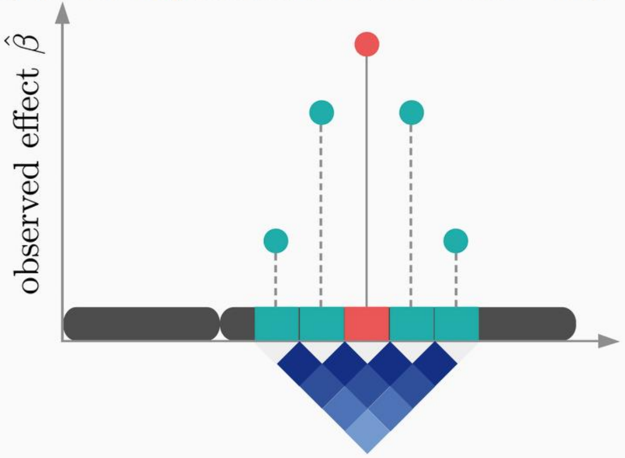


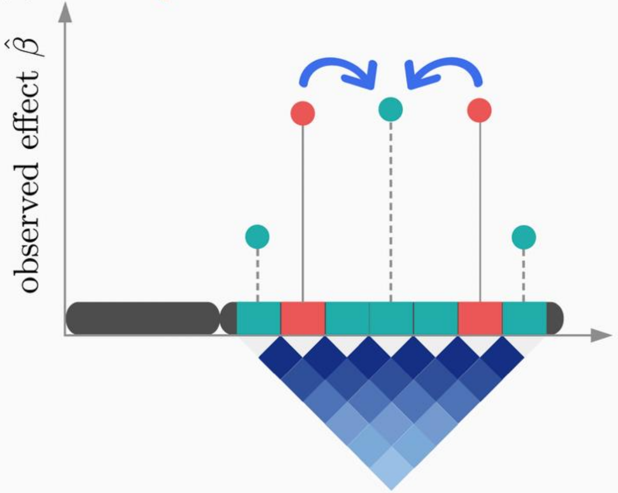
**Figure:** UK Biobank height GWAS, `http://nealelab.is/uk-biobank`



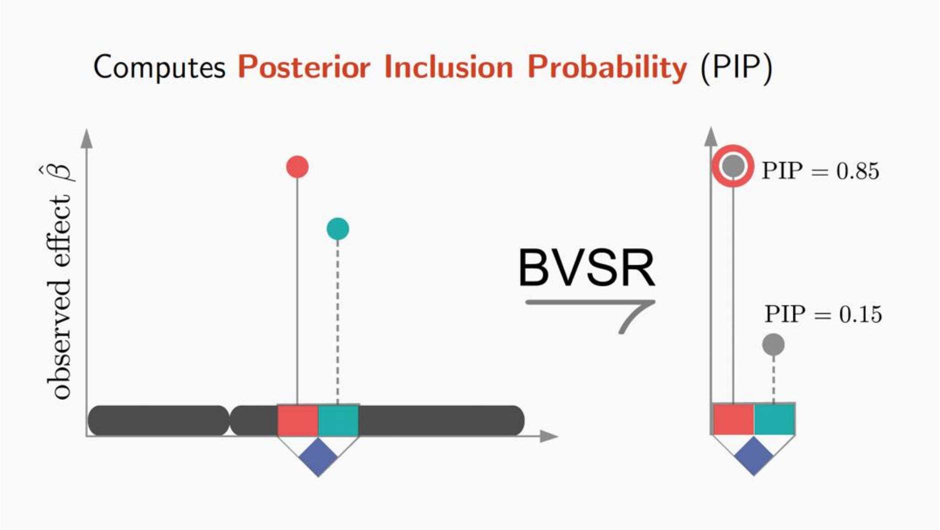
Simply pick the **top** association in an LD block? Maybe?

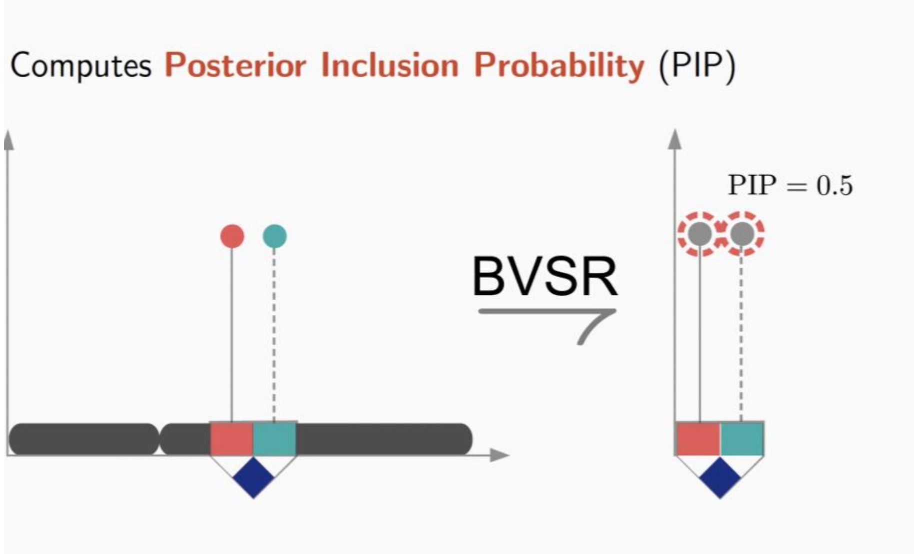


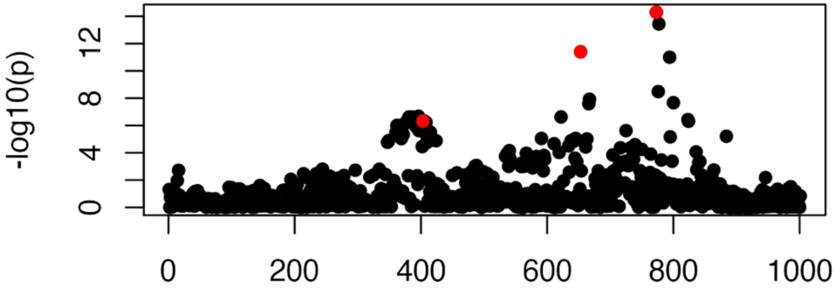
Simply pick the **top** association in an LD block? ... or not!

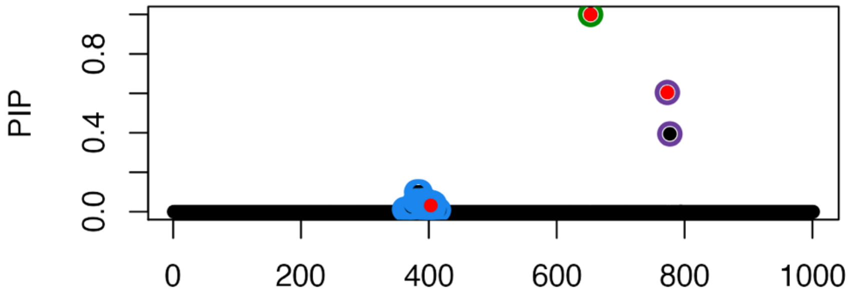


# SuSIE: Method to perform Bayesian variable selection to identify independent causal GWAS variants or sets of variants when it is not sure


Computes Posterior Inclusion Probability (PIP)
observed effect $\hat{\beta}$
BVSR
PIP = 0.85
PIP = 0.15



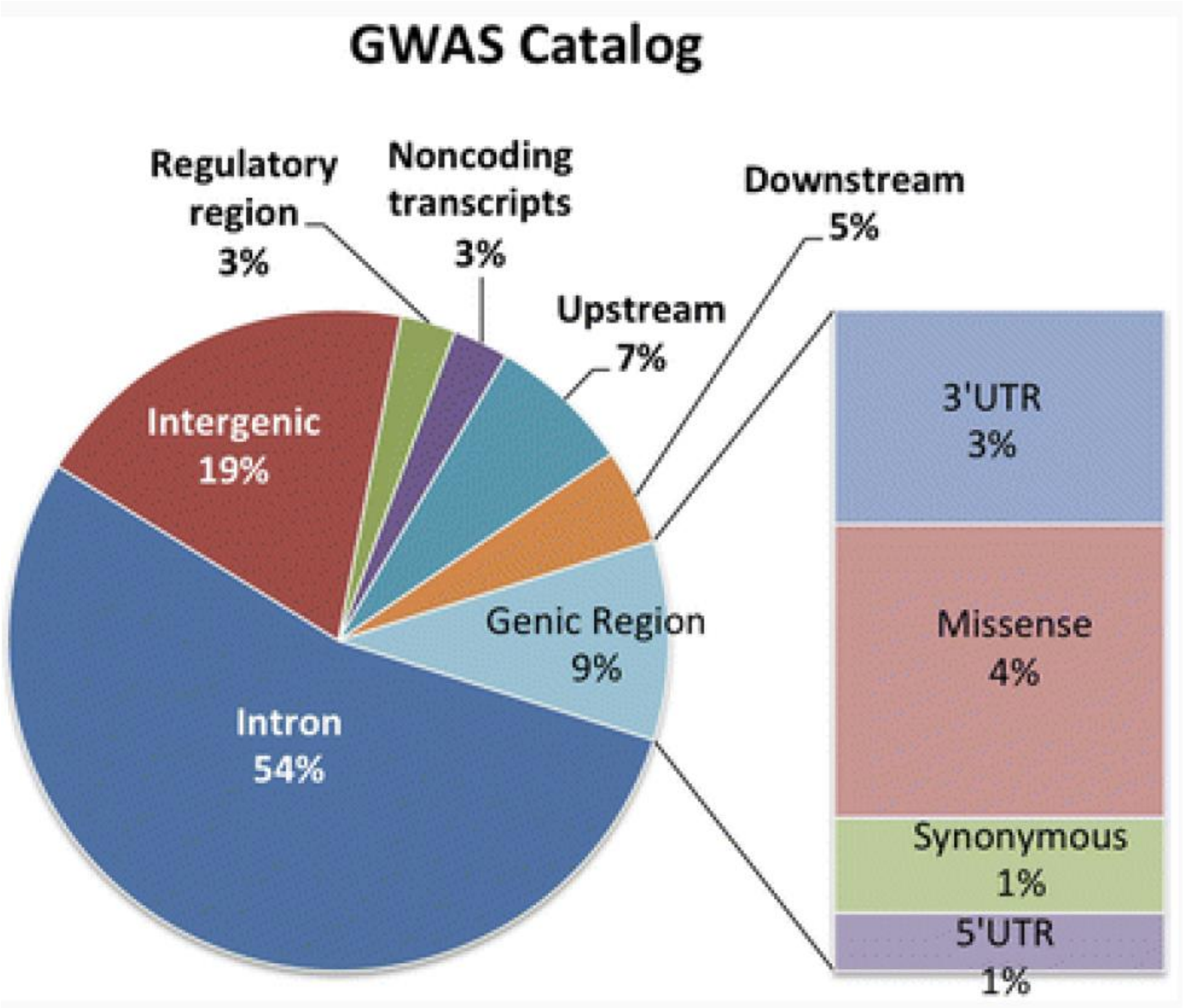




3 colors correspond to 95% **credible sets**: A credible set says a causal variant is within this set with 95% probability

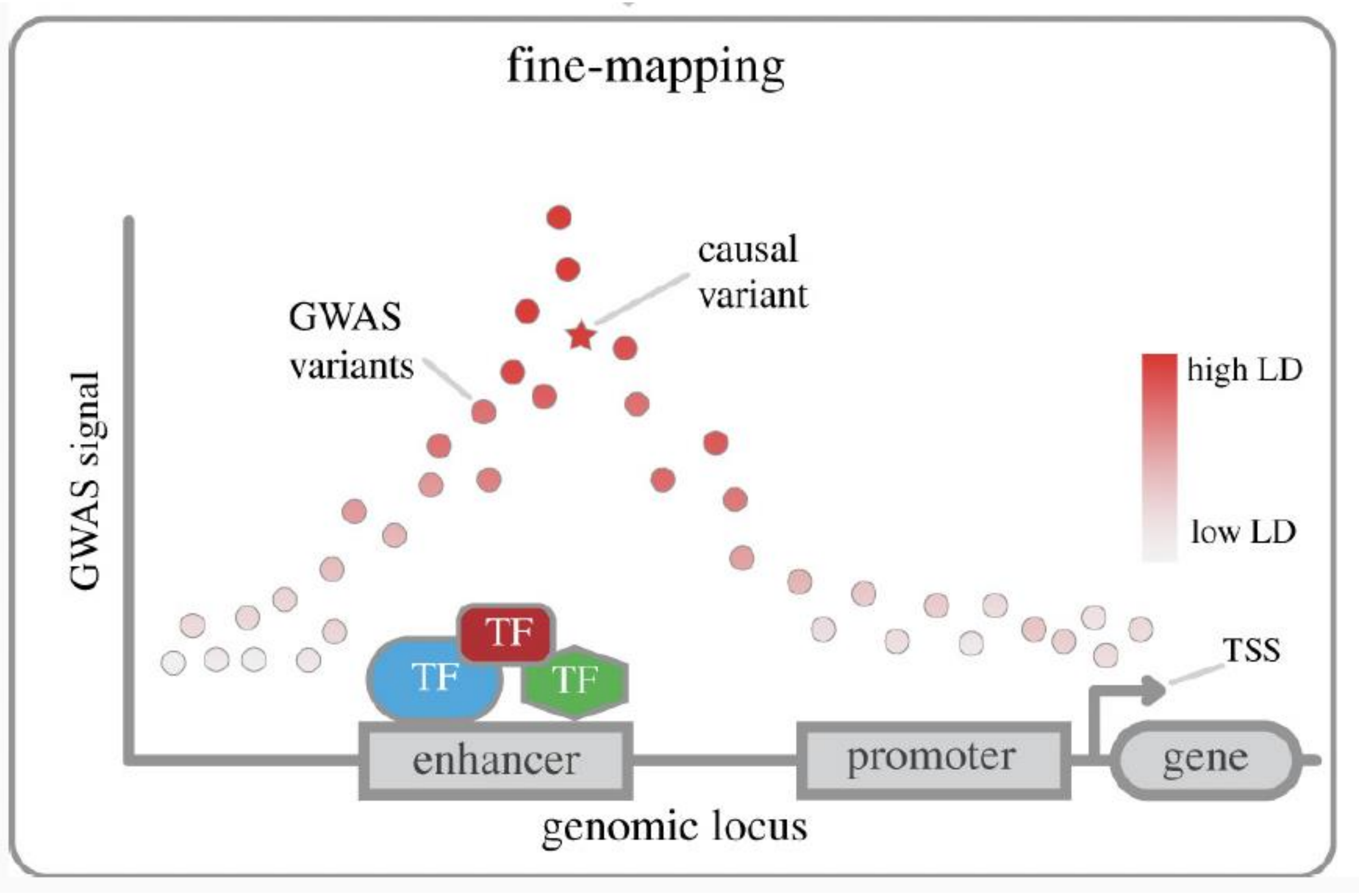
SuSIE: Wang et al 2020 *JRSS-B*

# Making sense of the function of GWAS variants

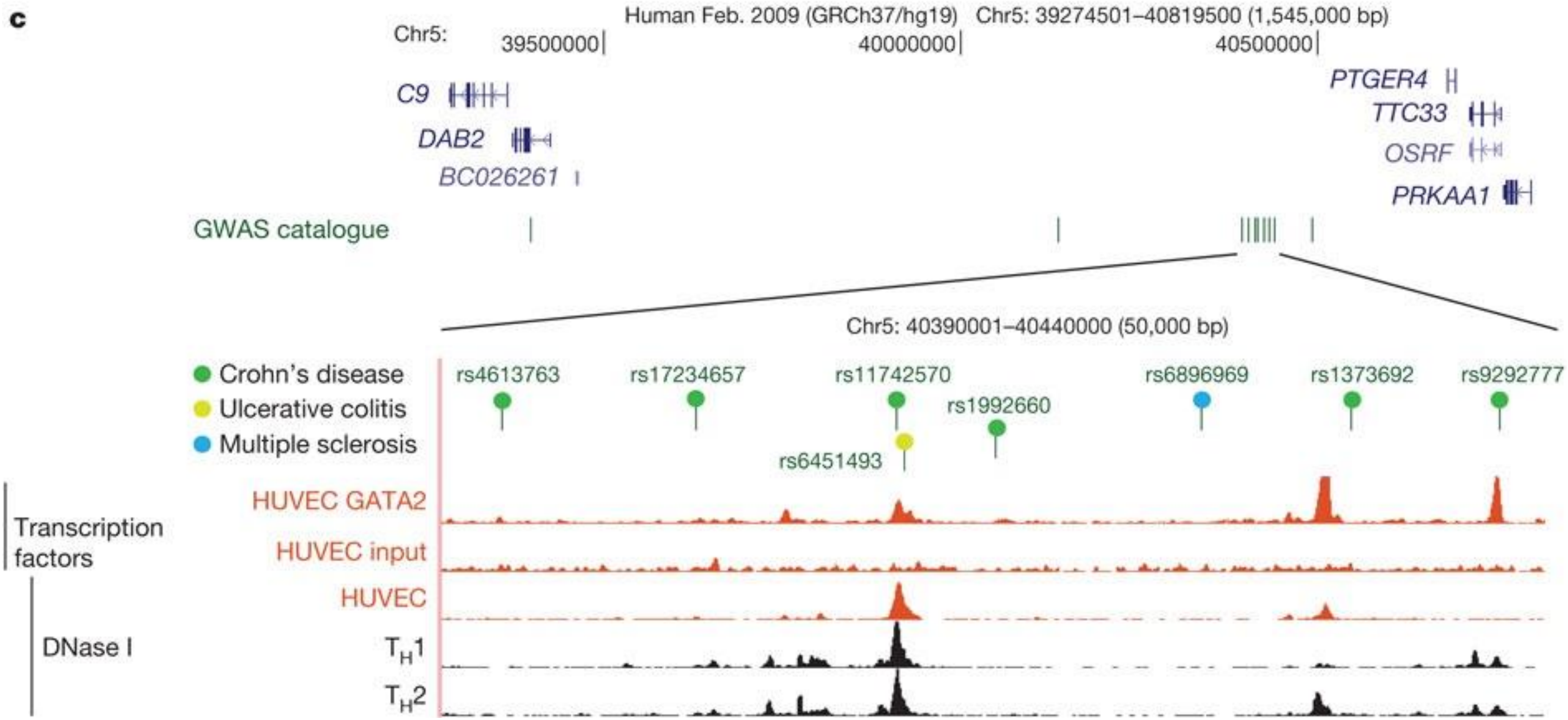


Lee et al 2018 *Human Genetics*

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



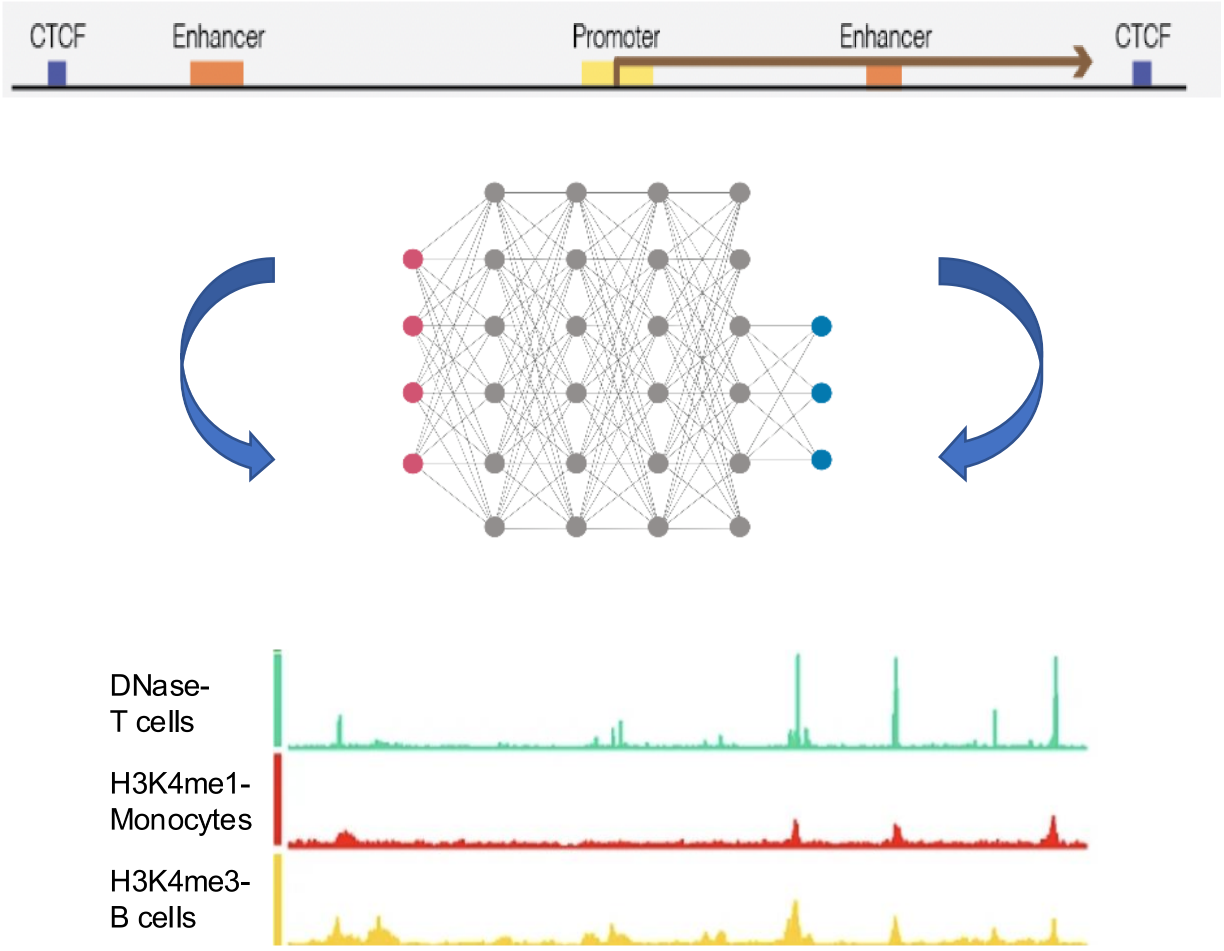
# Overlapping genome-wide functional annotation tracks against GWAS disease-associated variants



ENCODE consortium 2012, *Nature*

# Sequence-based deep learning models trained on epigenomic features

DNA sequences
(1-hot encoding DNA)

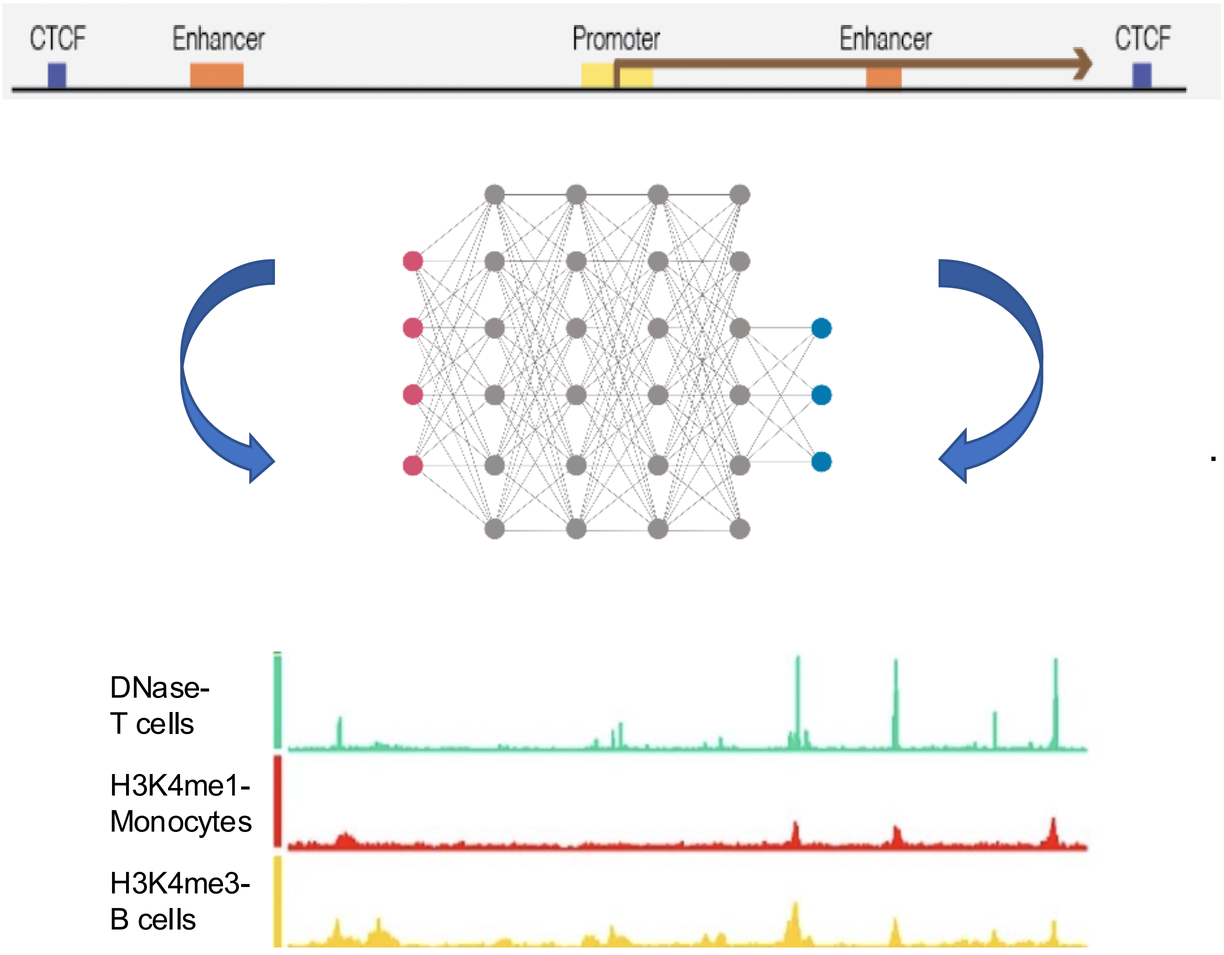


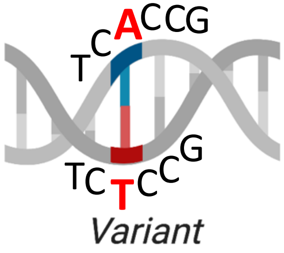
For each sequence, generates a prediction of affinity for each feature $f$ at the site of the sequence.

# Sequence-based deep learning models trained on epigenomic features

DNA sequences

(1-hot encoding DNA)



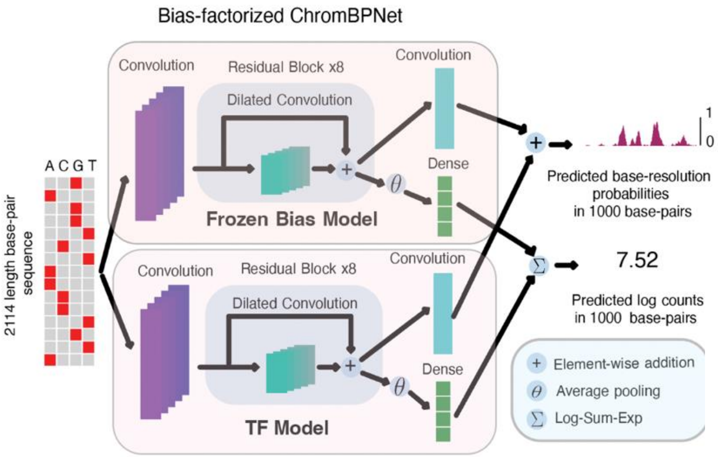


…TCACCG… …TCTCCG…

$p^f$ $q^f$

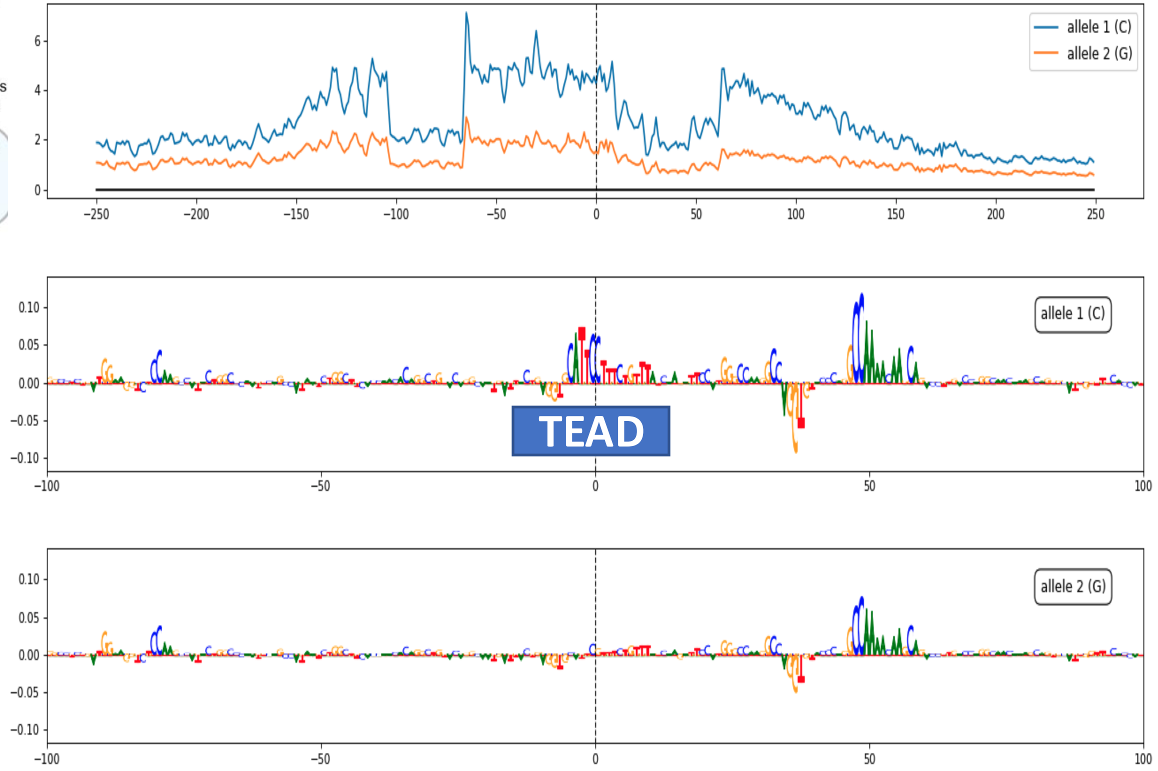
$$\Delta^f = | p^f - q^f |$$

# ChromBPNet deep learning model captures sequence mediated function at GWAS variants



Coronary Artery Disease GWAS variant

**rs4266144 (C/G): Human Quiescent SMCs**



https://github.com/kundajelab/chrombpnet

Also see

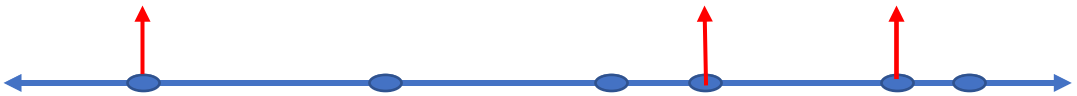
**Enformer**: Avsec et al 2021 *Nat Methods*
***BPNet***: : Avsec et al 2021 *Nat Genet*

Pampari et al 2024 bioRxiv
Courtesy: Anshul Kundaje, Stanford

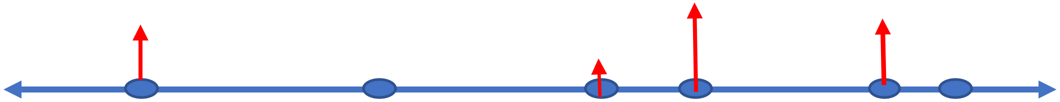
# Defining functional annotations at the level of variants

- Assigning a score to each SNP based on

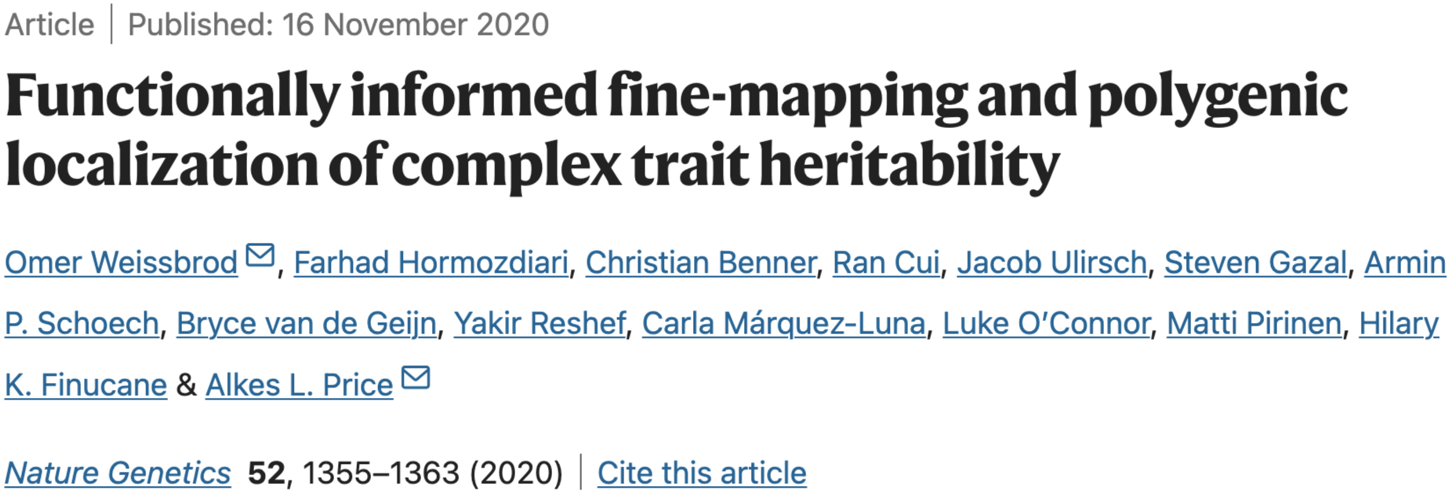
**Binary**: Presence or absence of a specific functional element at or around the SNP (example: SNP gets a score of 1 if there is a H3K4me1 peak at or around it)



**Continuous** value (often probabilistic scale between 0 and 1) measuring the strength of a specific function at or around the SNP (example: SNP is assigned the score equalling to the H3K4me1 peak intensity



# Using 97 functional annotations as prior improves the detection of causal variants

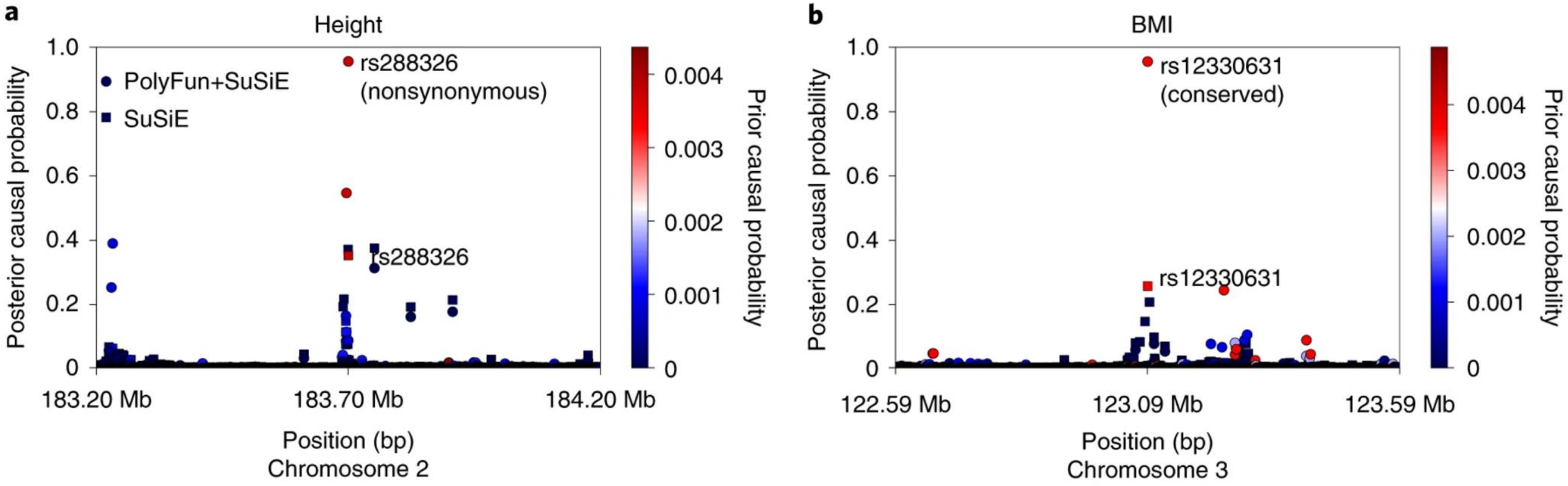


Article | Published: 16 November 2020

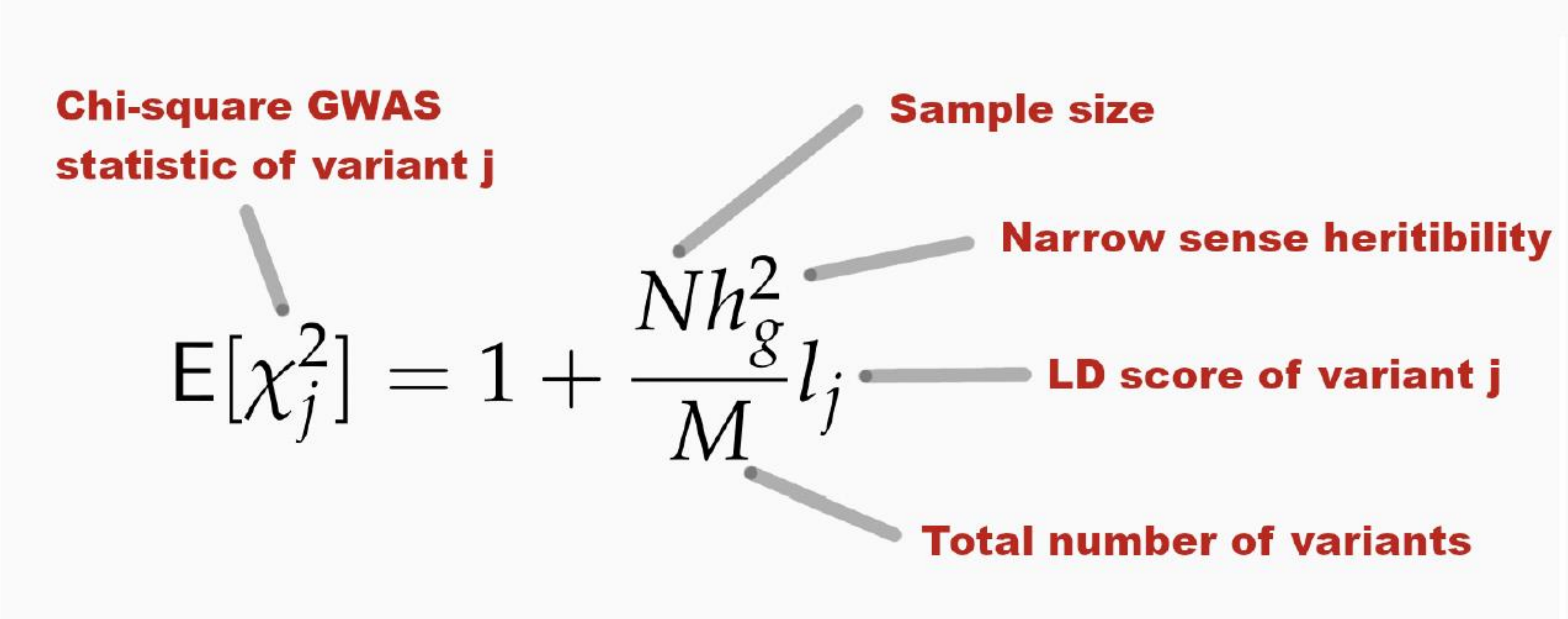
## Functionally informed fine-mapping and polygenic localization of complex trait heritability

Omer Weissbrod ✉, Farhad Hormozdiari, Christian Benner, Ran Cui, Jacob Ulirsch, Steven Gazal, Armin P. Schoech, Bryce van de Geijn, Yakir Reshef, Carla Márquez-Luna, Luke O'Connor, Matti Pirinen, Hilary K. Finucane & Alkes L. Price ✉

*Nature Genetics* **52**, 1355–1363 (2020) | Cite this article



# Mathematical overview of LD score regression



Chi-square GWAS statistic of variant j

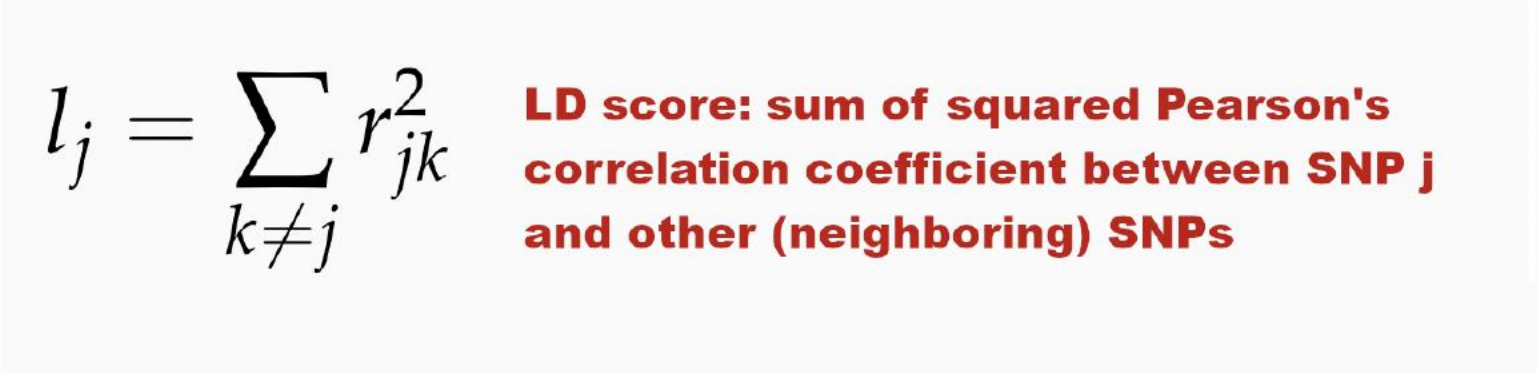
Sample size

Narrow sense heritibility

$$\mathrm{E}[\chi_j^2] = 1 + \frac{Nh_g^2}{M} l_j$$

LD score of variant j

Total number of variants



$$l_j = \sum_{k \neq j} r_{jk}^2$$

LD score: sum of squared Pearson's correlation coefficient between SNP j and other (neighboring) SNPs

Bulik-Sullivan Loh et al, 2015, *Nat Genet*

# Stratified LD score regression : Heritability enrichment due to functional categories of SNPs

Intuition: A category *f* is enriched for heritability if SNPs with high LD to that category have higher $\chi^2$ statistics.

$$\chi^2 \ = \ i \ + \sum_f N\tau_f \ \ LDscore_f$$

$LDscore_f \ (SNP\ x)$ = $\sum_{m \in f} r^2\ (x, m)$

Define heritability due to a functional category *f*

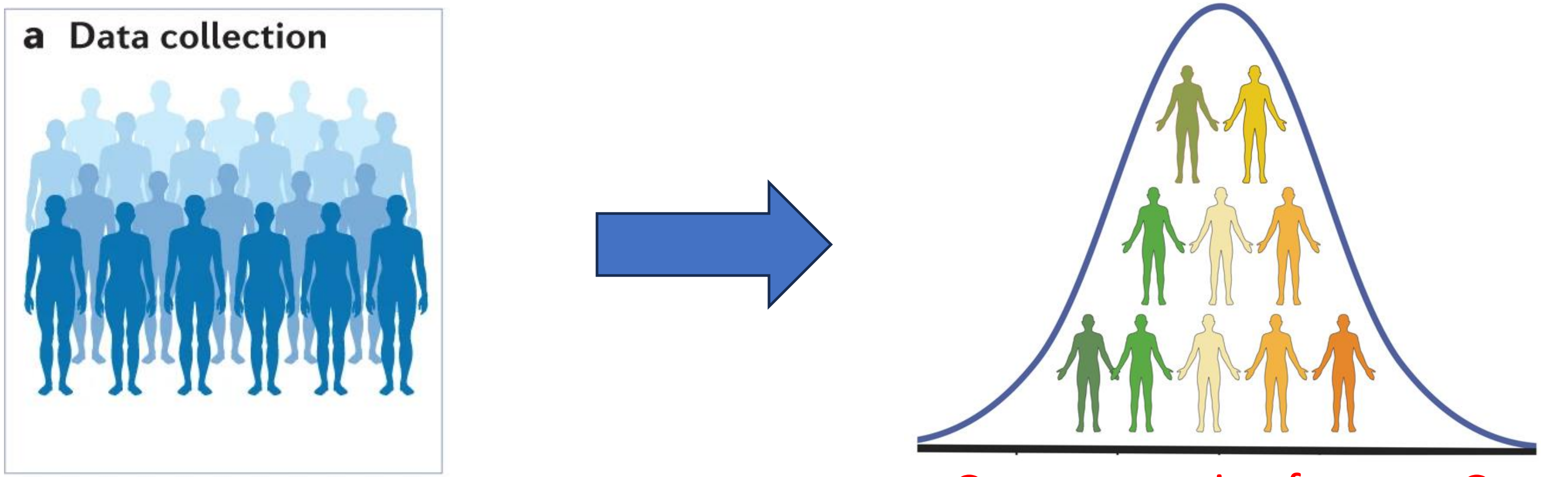
$$h_g^2(f) \ := \ \sum_{\{k \in f\}} \sum_{\{g\ contains\ k\}} \tau_g$$

Heritability enrichment (*f*) :=. $(h_g^2(f)/\ h_g^2\ )/(M(f)/M$

Finucane et al, 2015, *Nat Genet*

# Naturally occurring perturbations for human molecular phenotypes (QTLs)

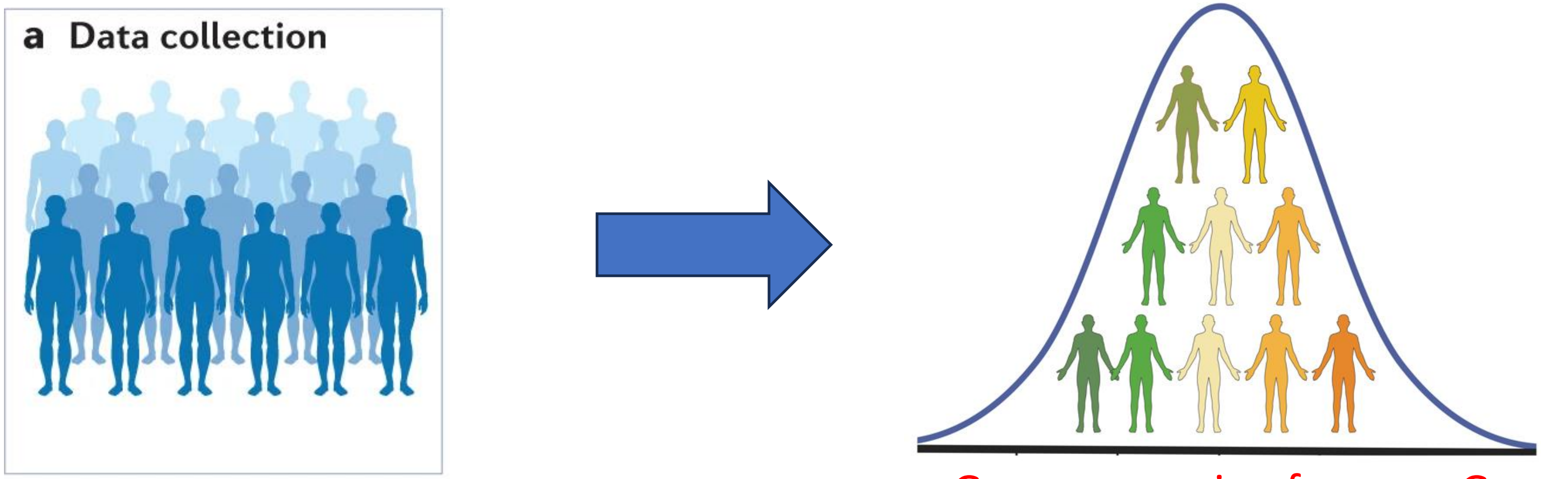
# Tracking genetic variation of gene expression phenotype (eQTL : expression quantitative trait loci)



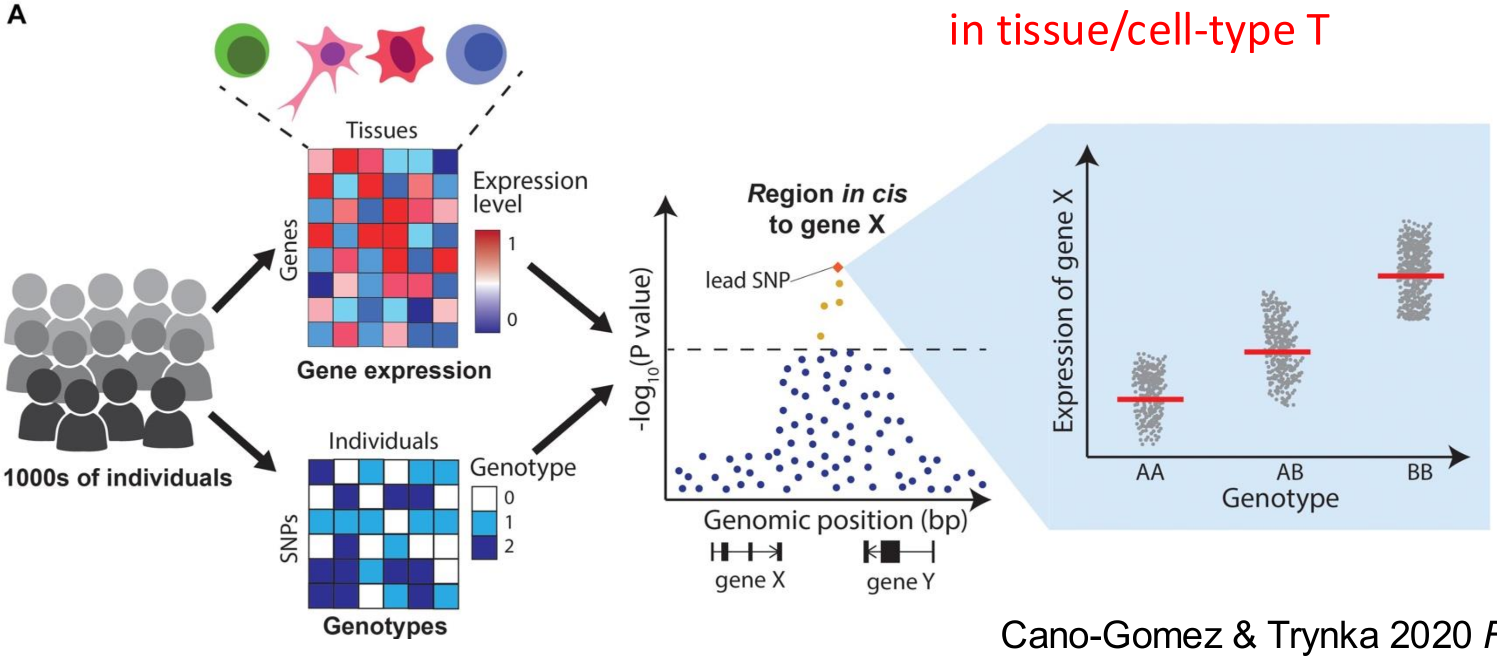
Gene expression for gene G
in tissue/cell-type T

Cano-Gomez & Trynka 2020 *Front Genet*

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



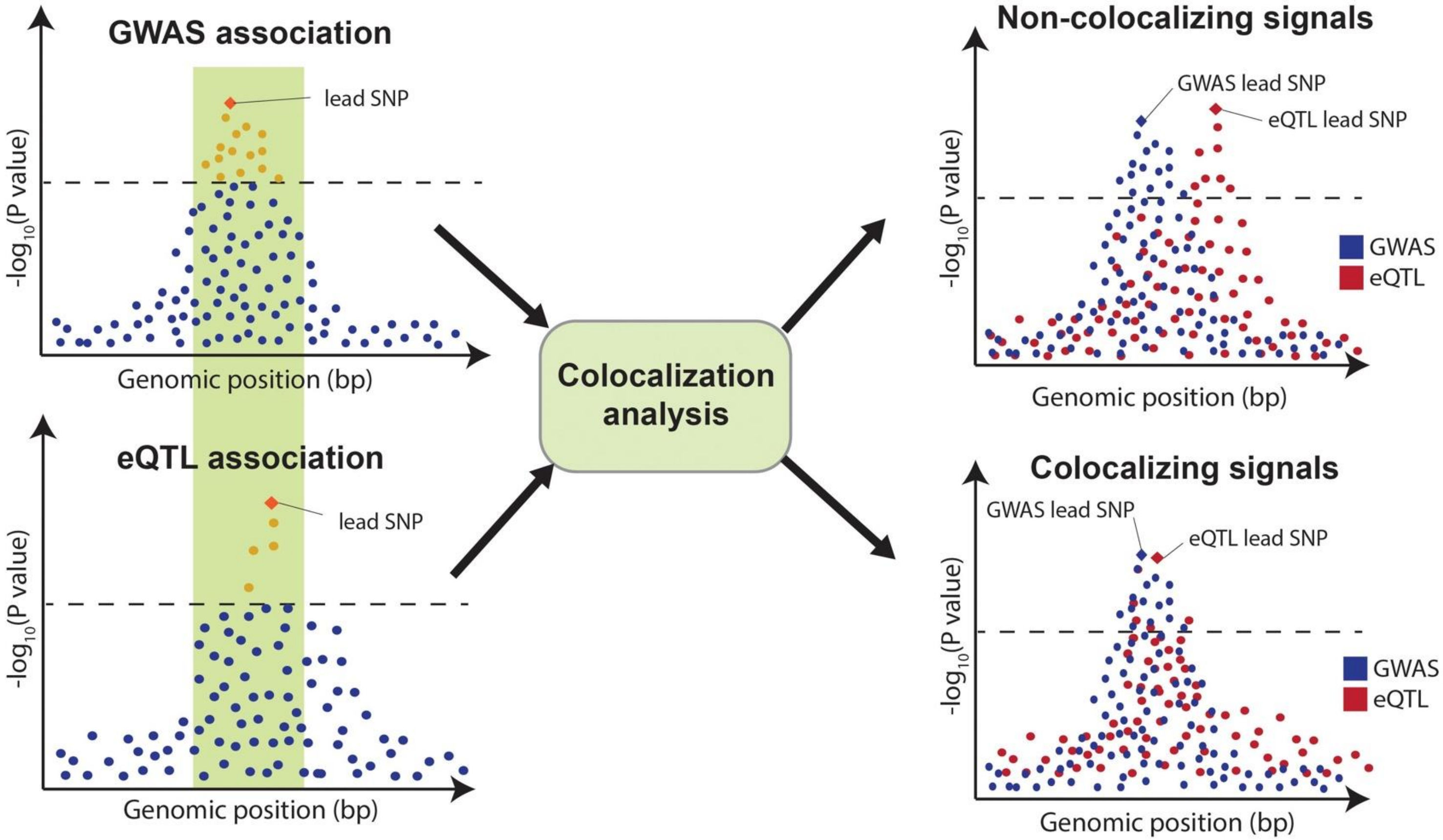
Gene expression for gene G in tissue/cell-type T



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.

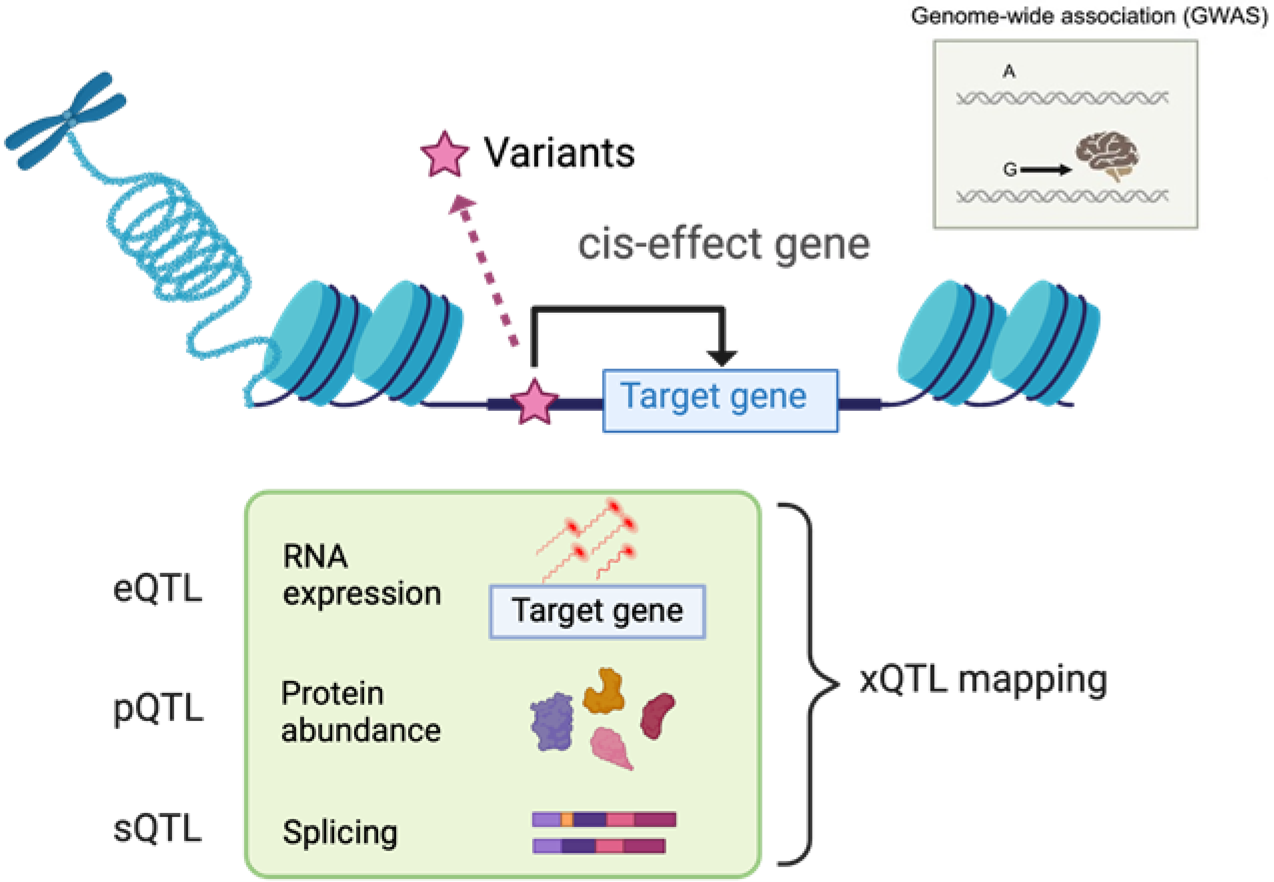


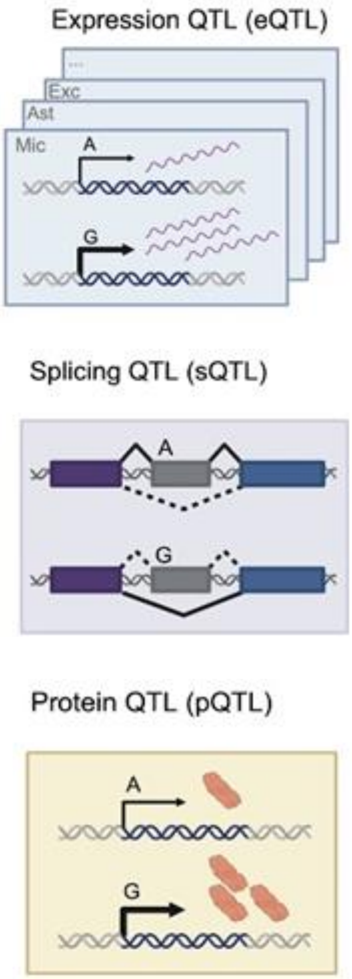
**Coloc**: standard method for colocalization.
Does not scale well to more than 2 phenotypes.

Giambartolomei et al 2014 *PLoS Gen*

# ColocBoost model to perform multimodal molecular phenotype QTL colocalization

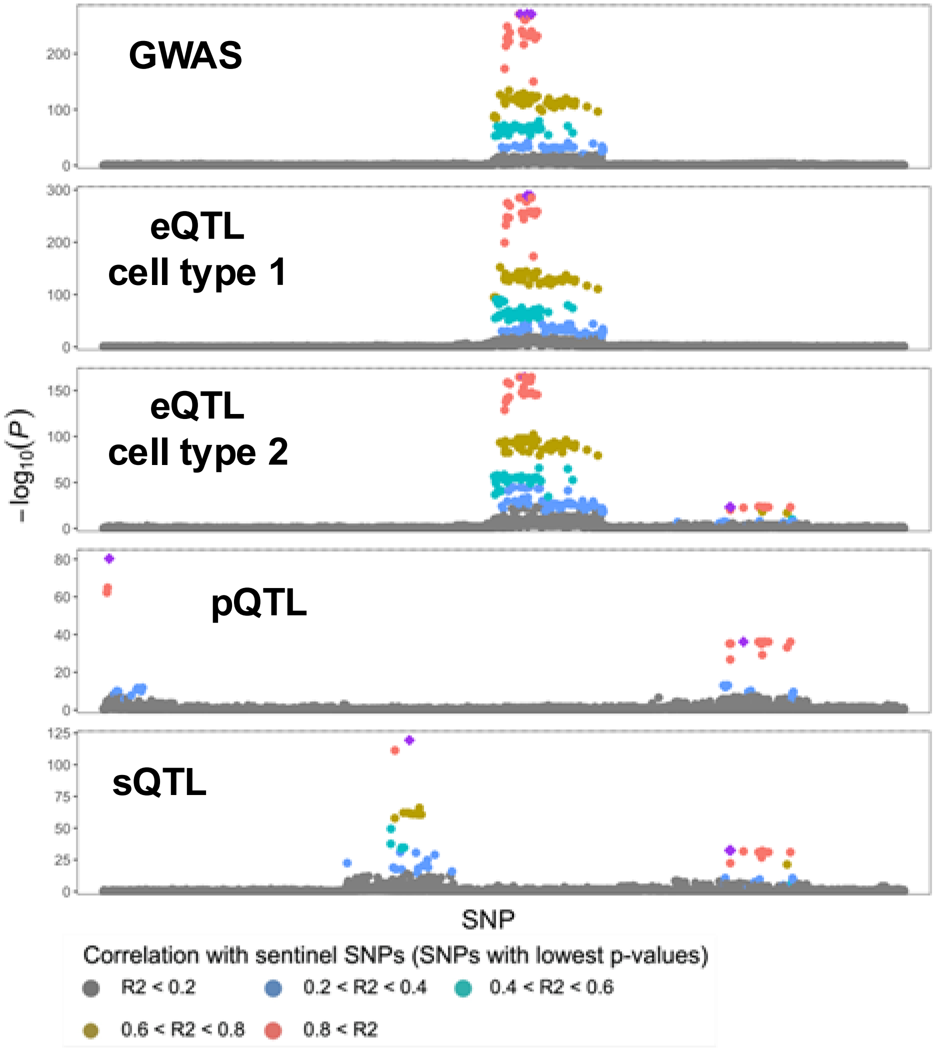
Recent advances in technology has made it easier to use other molecular phenotypes outside of gene expression, and also assess eQTL at cell type resolution for different cell types in a tissue

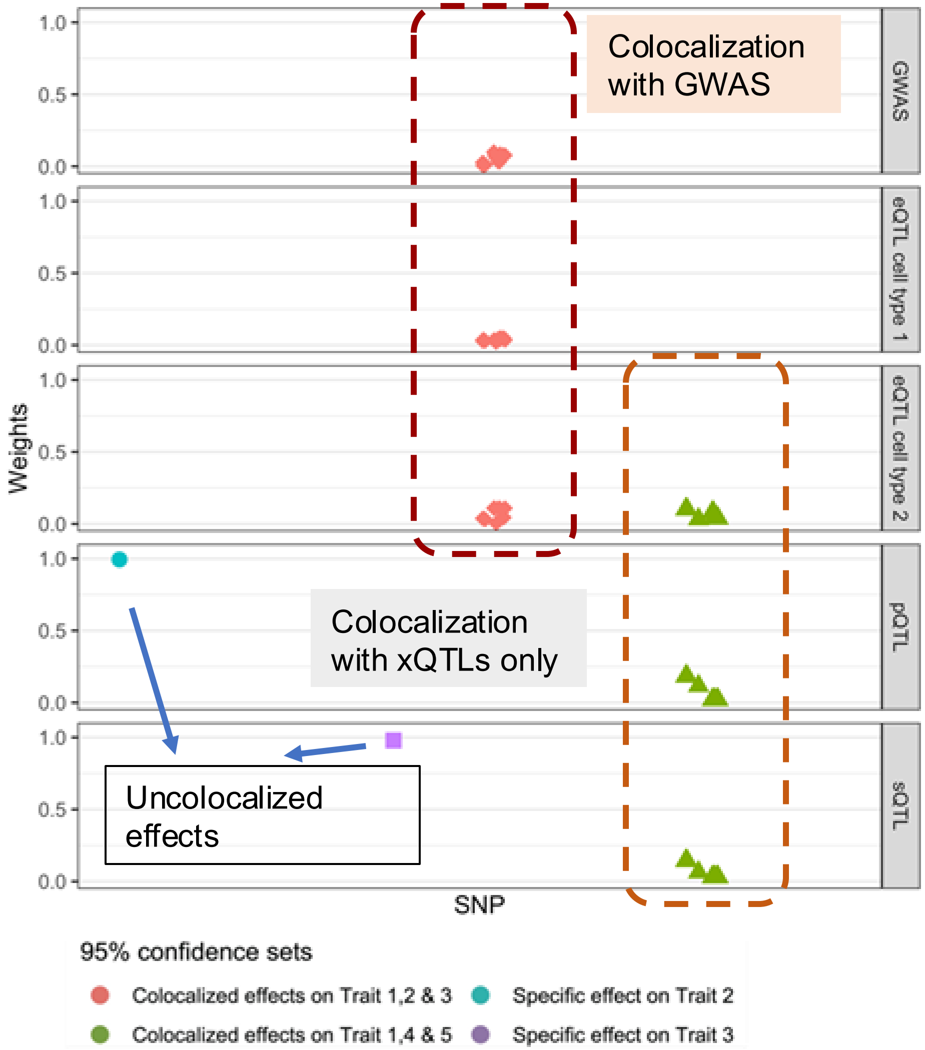




Aguet et al. 2020. *Nature Reviews Methods Primers*

# Understanding colocalization: enhancing GWAS insights through shared genetic signals





Colocalization with GWAS

Colocalization with xQTLs only

Uncolocalized effects

Memorial Sloan Kettering Cancer Center

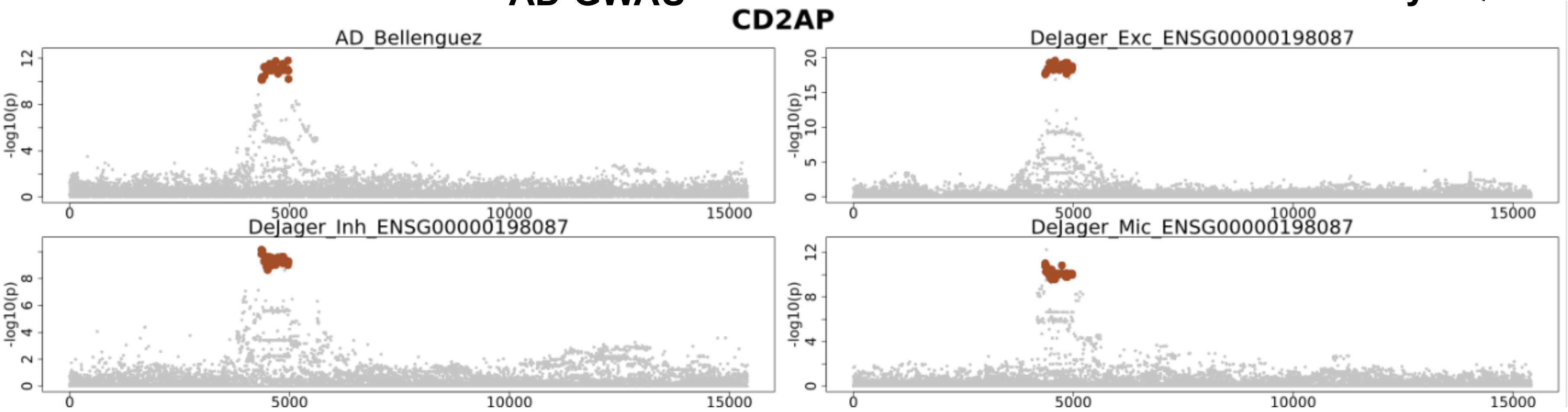
ColocBoost: Cao et al 2025 medRxiv, in rev *Nat Genet*
HyPrColoc: Foley et al 2021 *Nat Commun*

# 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.

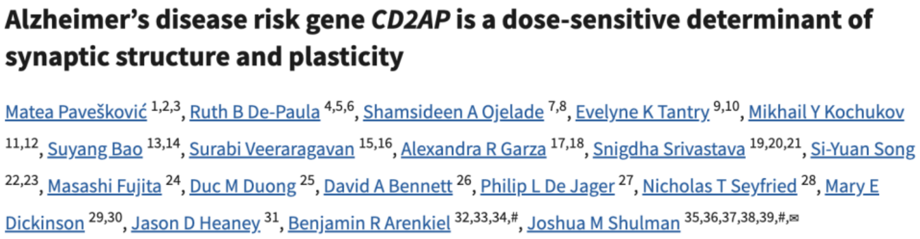
AD GWAS

Excitatory eQTL



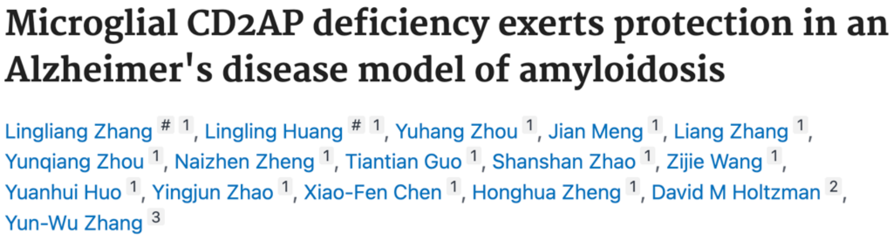
Inhibitory eQTL

Microglia eQTL



**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*

# 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

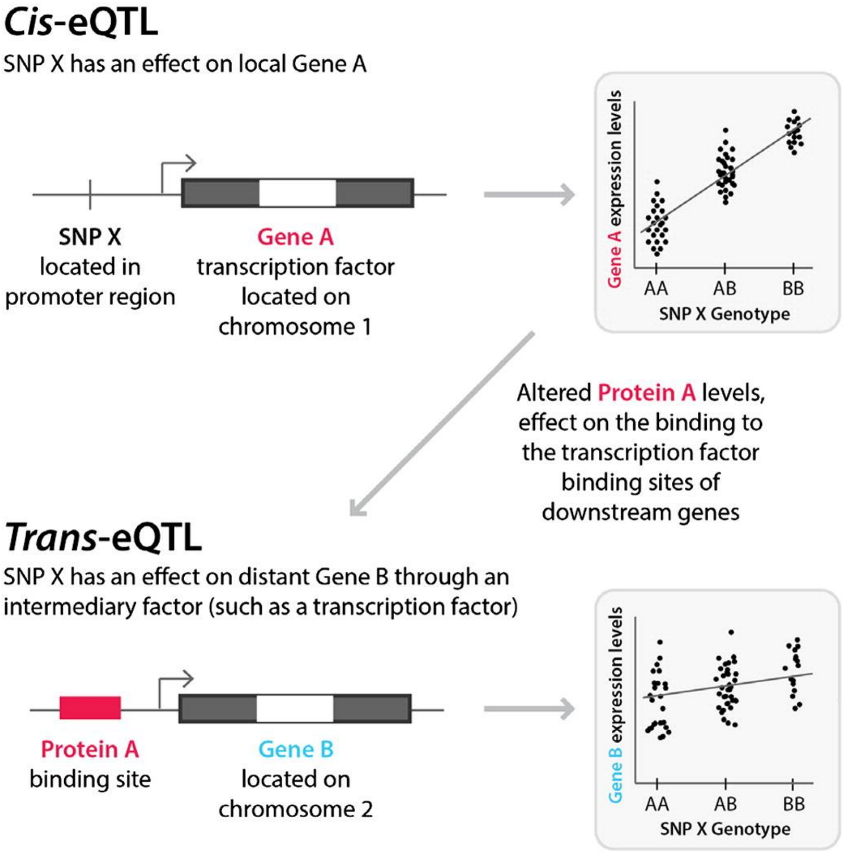
Hakhamanesh Mostafavi [1], Jeffrey P Spence [2], Sahin Naqvi [2] [3], Jonathan K Pritchard [4] [5]

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****, show relaxed constraint, and have simpler regulatory landscapes. ”*

# Cis and trans-eQTLs can identify proximal and distal genes of action



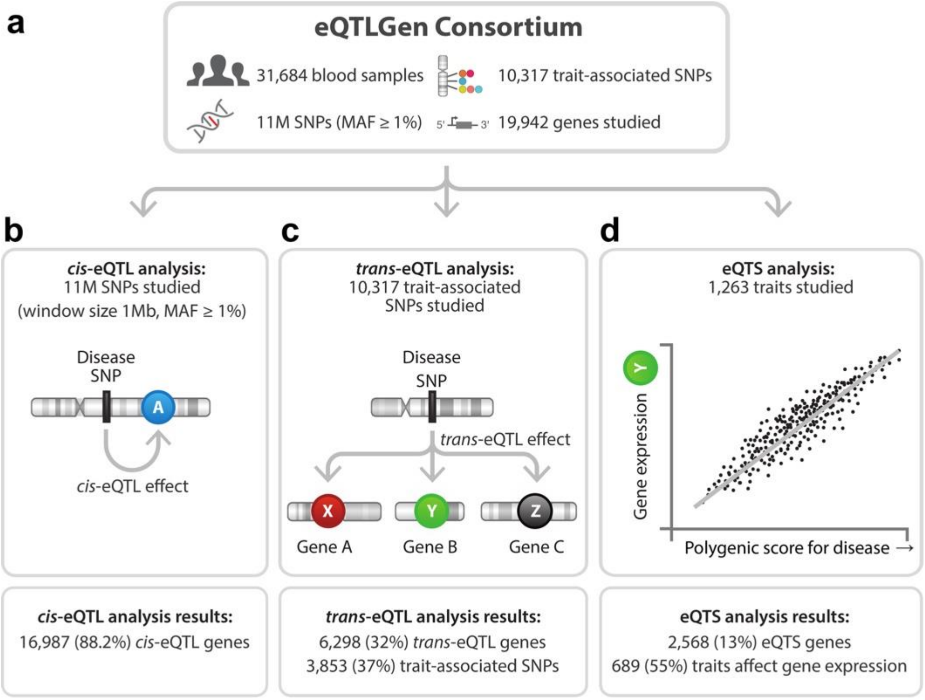
***Cis*-eQTL**

SNP X has an effect on local Gene A

Altered Protein A levels, effect on the binding to the transcription factor binding sites of downstream genes

***Trans*-eQTL**

SNP X has an effect on distant Gene B through an intermediary factor (such as a transcription factor)



***cis*-eQTL analysis:**
11M SNPs studied
(window size 1Mb, MAF ≥ 1%)

***trans*-eQTL analysis:**
10,317 trait-associated
SNPs studied

**eQTS analysis:**
1,263 traits studied

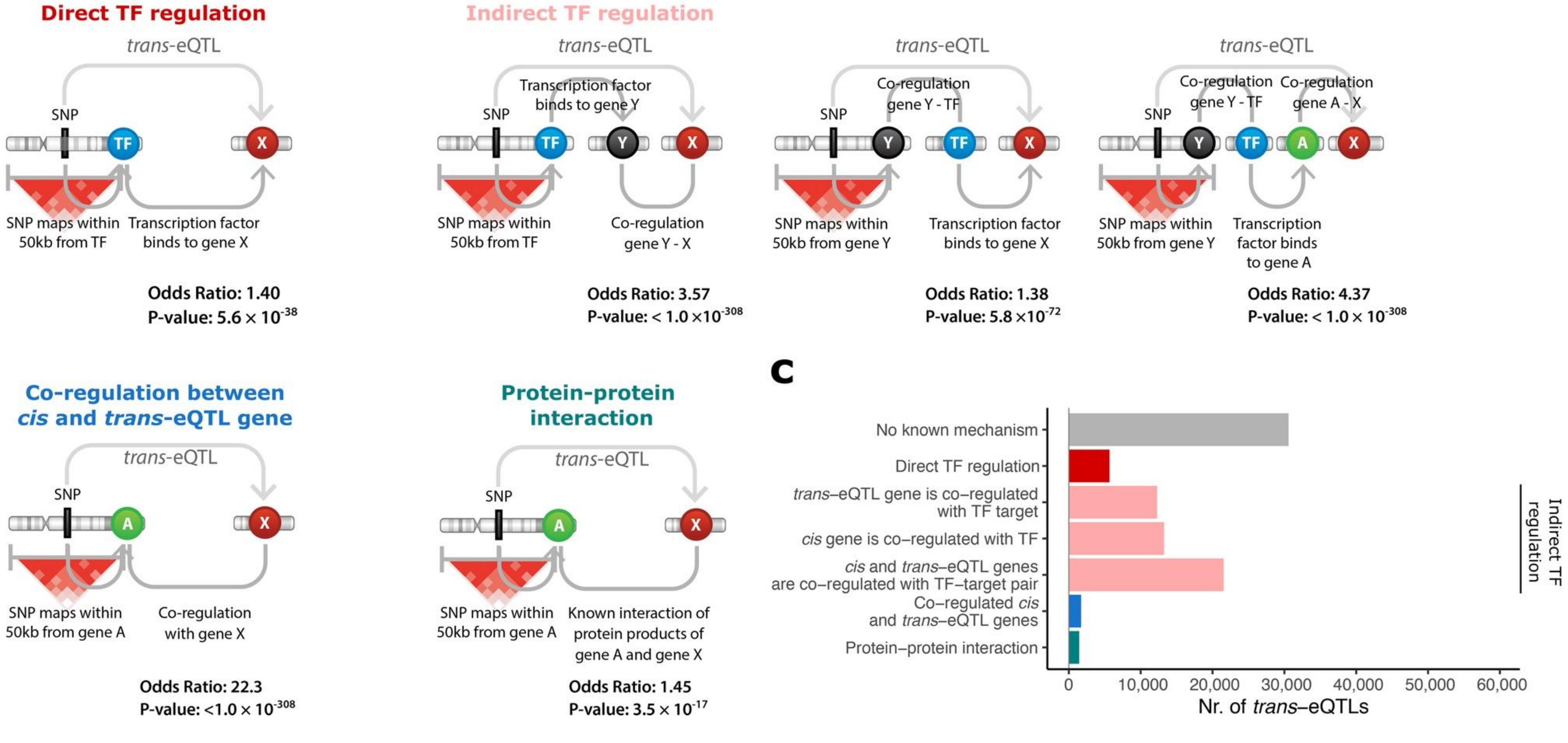
***cis*-eQTL analysis results:**
16,987 (88.2%) *cis*-eQTL genes

***trans*-eQTL analysis results:**
6,298 (32%) *trans*-eQTL genes
3,853 (37%) trait-associated SNPs

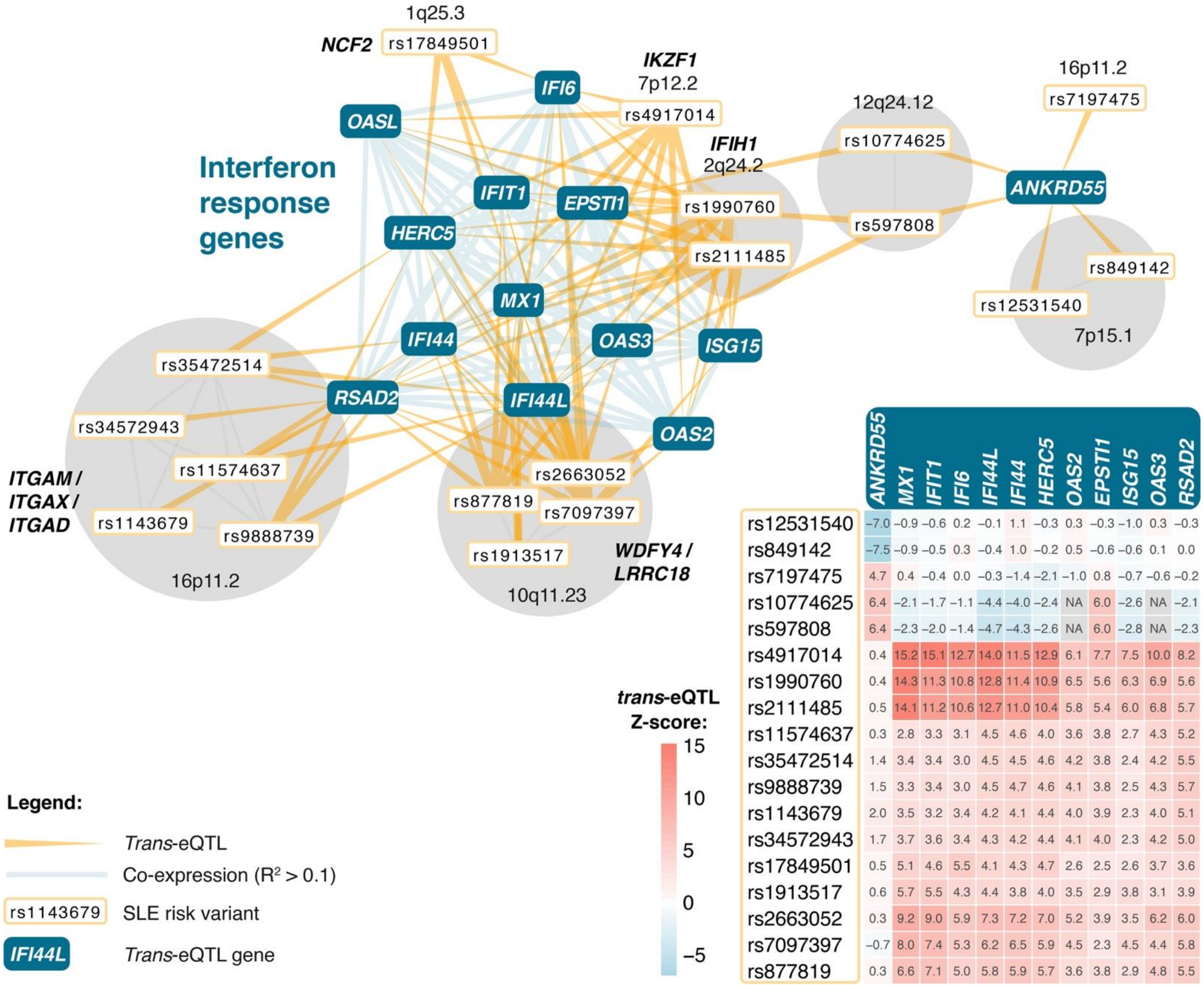
**eQTS analysis results:**
2,568 (13%) eQTS genes
689 (55%) traits affect gene expression

Westra and Franke 2014 *BBA*.
Vosa et al 2021 *Nat Genet*

# Cis and trans-eQTLs can identify proximal and distal genes of action

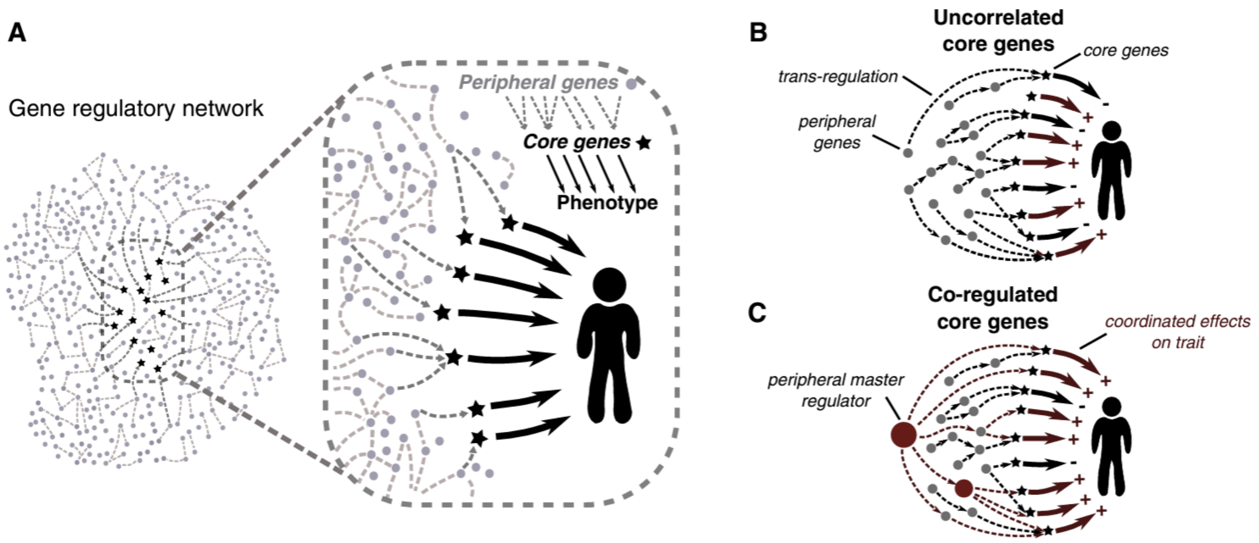

Direct TF regulation
trans-eQTL
SNP
TF
X
SNP maps within 50kb from TF
Transcription factor binds to gene X
Odds Ratio: 1.40
P-value: 5.6 × 10-38
Indirect TF regulation
trans-eQTL
Transcription factor binds to gene Y
SNP
TF
Y
X
SNP maps within 50kb from TF
Co-regulation gene Y - X
Odds Ratio: 3.57
P-value: < 1.0 ×10-308
trans-eQTL
Co-regulation gene Y - TF
SNP
Y
TF
X
SNP maps within 50kb from gene Y
Transcription factor binds to gene X
Odds Ratio: 1.38
P-value: 5.8 ×10-72
trans-eQTL
Co-regulation gene Y - TF
Co-regulation gene A - X
SNP
Y
TF
A
X
SNP maps within 50kb from gene Y
Transcription factor binds to gene A
Odds Ratio: 4.37
P-value: < 1.0 × 10-308
Co-regulation between cis and trans-eQTL gene
trans-eQTL
SNP
A
X
SNP maps within 50kb from gene A
Co-regulation with gene X
Odds Ratio: 22.3
P-value: <1.0 × 10-308
Protein-protein interaction
trans-eQTL
SNP
A
X
SNP maps within 50kb from gene A
Known interaction of protein products of gene A and gene X
Odds Ratio: 1.45
P-value: 3.5 × 10-17
c
No known mechanism
Direct TF regulation
trans–eQTL gene is co–regulated with TF target
cis gene is co–regulated with TF
cis and trans–eQTL genes are co–regulated with TF–target pair
Co–regulated cis and trans–eQTL genes
Protein–protein interaction
Indirect TF regulation
0
10,000
20,000
30,000
40,000
50,000
60,000
Nr. of trans–eQTLs

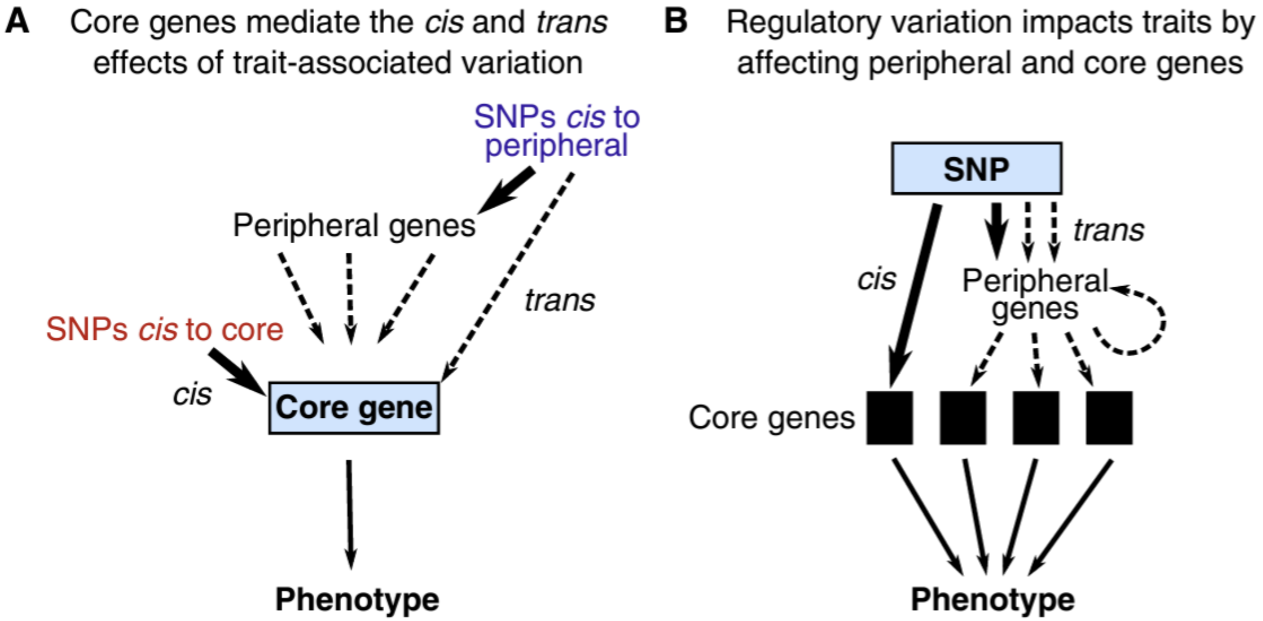
# Cis and trans-eQTLs can identify proximal and distal genes of action



| | *ANKRD55* | *MX1* | *IFIT1* | *IFI6* | *IFI44L* | *IFI44* | *HERC5* | *OAS2* | *EPSTI1* | *ISG15* | *OAS3* | *RSAD2* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs12531540 | −7.0 | −0.9 | −0.6 | 0.2 | −0.1 | 1.1 | −0.3 | 0.3 | −0.3 | −1.0 | 0.3 | −0.3 |
| rs849142 | −7.5 | −0.9 | −0.5 | 0.3 | −0.4 | 1.0 | −0.2 | 0.5 | −0.6 | −0.6 | 0.1 | 0.0 |
| rs7197475 | 4.7 | 0.4 | −0.4 | 0.0 | −0.3 | −1.4 | −2.1 | −1.0 | 0.8 | −0.7 | −0.6 | −0.2 |
| rs10774625 | 6.4 | −2.1 | −1.7 | −1.1 | −4.4 | −4.0 | −2.4 | NA | 6.0 | −2.6 | NA | −2.1 |
| rs597808 | 6.4 | −2.3 | −2.0 | −1.4 | −4.7 | −4.3 | −2.6 | NA | 6.0 | −2.8 | NA | −2.3 |
| rs4917014 | 0.4 | 15.2 | 15.1 | 12.7 | 14.0 | 11.5 | 12.9 | 6.1 | 7.7 | 7.5 | 10.0 | 8.2 |
| rs1990760 | 0.4 | 14.3 | 11.3 | 10.8 | 12.8 | 11.4 | 10.9 | 6.5 | 5.6 | 6.3 | 6.9 | 5.6 |
| rs2111485 | 0.5 | 14.1 | 11.2 | 10.6 | 12.7 | 11.0 | 10.4 | 5.8 | 5.4 | 6.0 | 6.8 | 5.7 |
| rs11574637 | 0.3 | 2.8 | 3.3 | 3.1 | 4.5 | 4.6 | 4.0 | 3.6 | 3.8 | 2.7 | 4.3 | 5.2 |
| rs35472514 | 1.4 | 3.4 | 3.4 | 3.0 | 4.5 | 4.5 | 4.6 | 4.2 | 3.8 | 2.4 | 4.2 | 5.5 |
| rs9888739 | 1.5 | 3.3 | 3.4 | 3.0 | 4.5 | 4.7 | 4.6 | 4.1 | 3.8 | 2.5 | 4.3 | 5.7 |
| rs1143679 | 2.0 | 3.5 | 3.2 | 3.4 | 4.2 | 4.1 | 4.4 | 4.0 | 3.9 | 2.3 | 4.0 | 5.1 |
| rs34572943 | 1.7 | 3.7 | 3.6 | 3.4 | 4.3 | 4.2 | 4.4 | 4.1 | 4.0 | 2.3 | 4.2 | 5.0 |
| rs17849501 | 0.5 | 5.1 | 4.6 | 5.5 | 4.1 | 4.3 | 4.7 | 2.6 | 2.5 | 2.6 | 3.7 | 3.6 |
| rs1913517 | 0.6 | 5.7 | 5.5 | 4.3 | 4.4 | 3.8 | 4.0 | 3.5 | 2.9 | 3.8 | 3.1 | 3.9 |
| rs2663052 | 0.3 | 9.2 | 9.0 | 5.9 | 7.3 | 7.2 | 7.0 | 5.2 | 3.9 | 3.5 | 6.2 | 6.0 |
| rs7097397 | −0.7 | 8.0 | 7.4 | 5.3 | 6.2 | 6.5 | 5.9 | 4.5 | 2.3 | 4.5 | 4.4 | 5.8 |
| rs877819 | 0.3 | 6.6 | 7.1 | 5.0 | 5.8 | 5.9 | 5.7 | 3.6 | 3.8 | 2.9 | 4.8 | 5.5 |

# Omnigenic model hypothesis in genetics

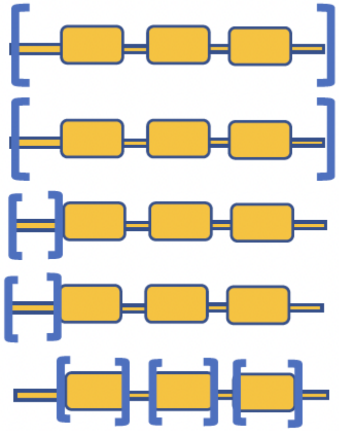



A Core genes mediate the *cis* and *trans* effects of trait-associated variation
SNPs *cis* to peripheral
Peripheral genes
*trans*
SNPs *cis* to core
*cis*
Core gene
Phenotype
B Regulatory variation impacts traits by affecting peripheral and core genes
SNP
*cis*
*trans*
Peripheral genes
Core genes
Phenotype

# Other approaches of mapping GWAS

# Variants to Genes (V2G)

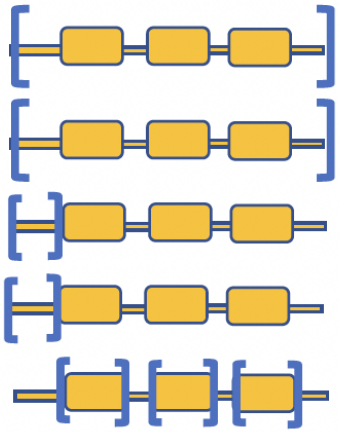
# Broadening the scope of approaches to link variants to genes



| 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*

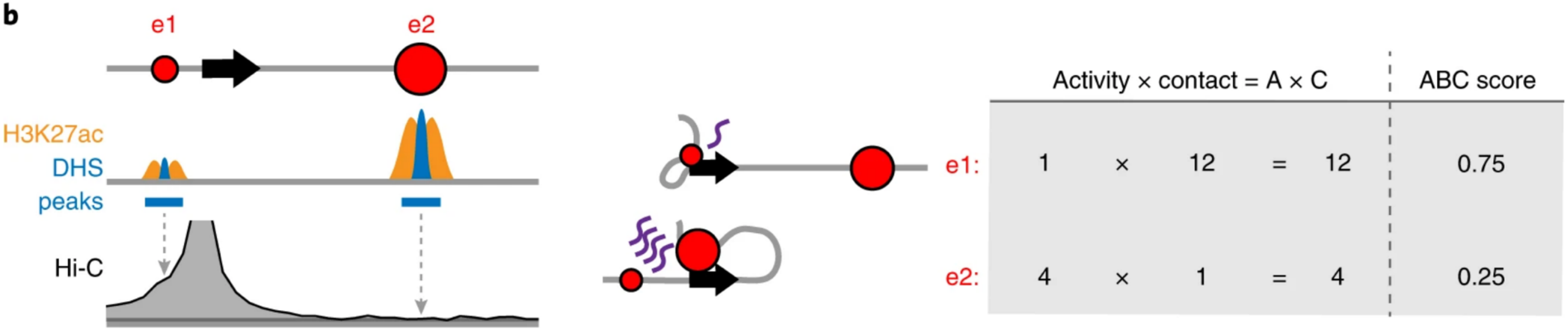
# Broadening the scope of approaches to link variants to genes



| 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 |

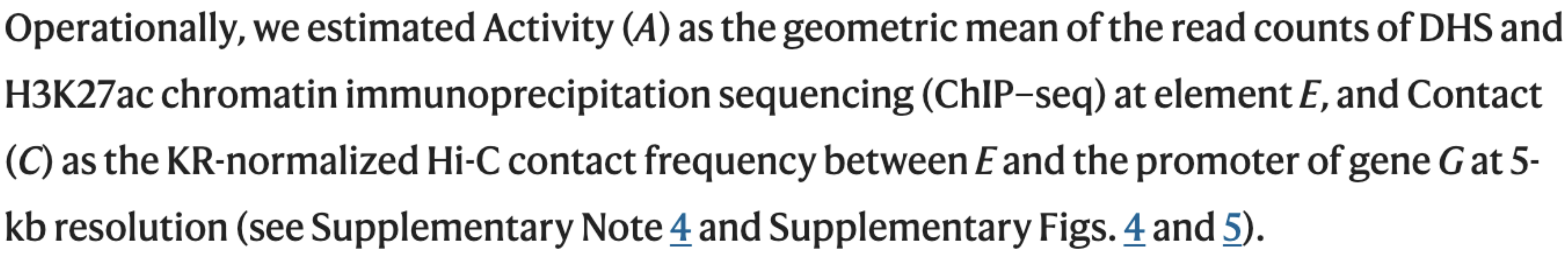
Dey et al 2022, *Cell Genomics,*
Gazal..Dey et al 2022 *Nat Genet*

# The Activity-By-Contact element-gene linking method



| | Activity | × | contact | = | A × C | ABC score |
|---|---|---|---|---|---|---|
| e1: | 1 | × | 12 | = | 12 | 0.75 |
| e2: | 4 | × | 1 | = | 4 | 0.25 |

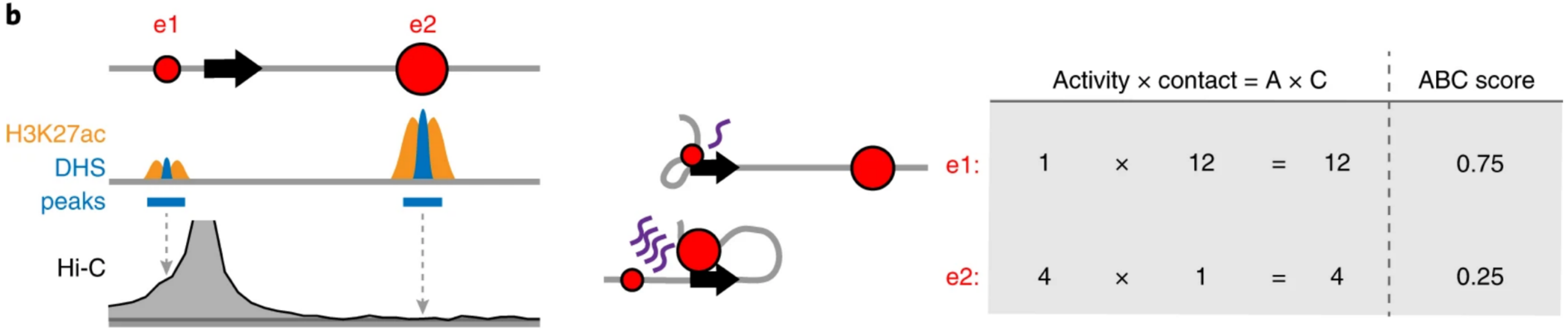
$$\text{ABC score}_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{\text{all elements e within 5 Mb of } G} A_e \times C_{e,G}}$$



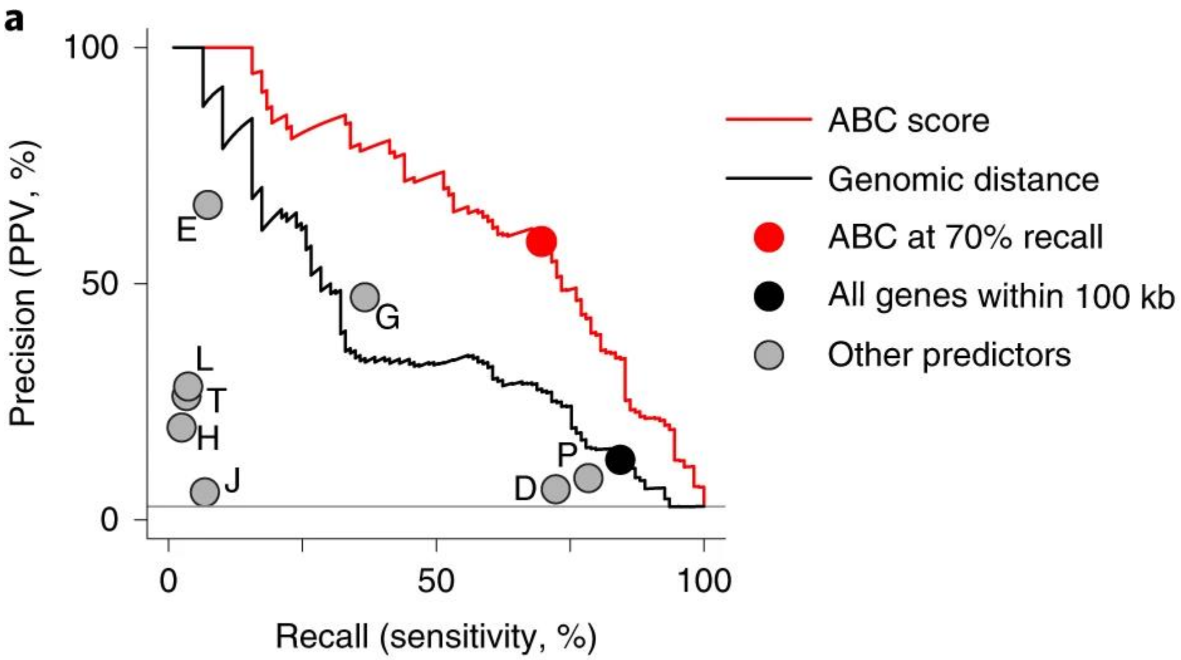
Operationally, we estimated Activity (*A*) as the geometric mean of the read counts of DHS and H3K27ac chromatin immunoprecipitation sequencing (ChIP–seq) at element *E*, and Contact (*C*) as the KR-normalized Hi-C contact frequency between *E* and the promoter of gene *G* at 5-kb resolution (see Supplementary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

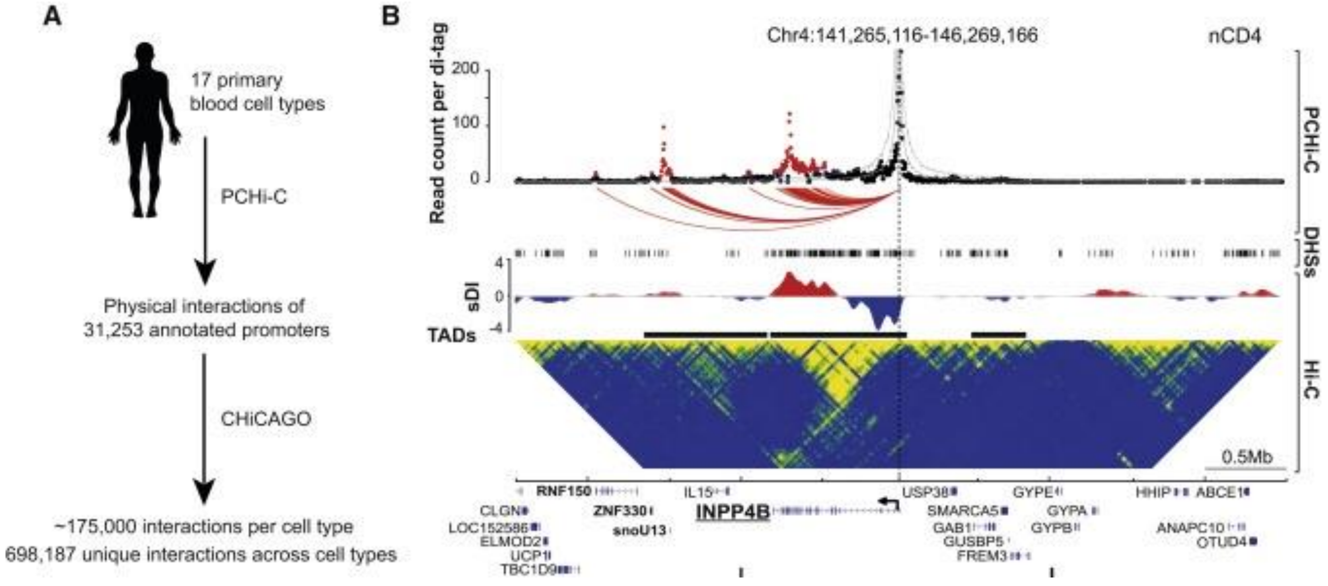


| | Activity × contact = A × C | | | | | ABC score |
|---|---|---|---|---|---|---|
| e1: | 1 | × | 12 | = | 12 | 0.75 |
| e2: | 4 | × | 1 | = | 4 | 0.25 |

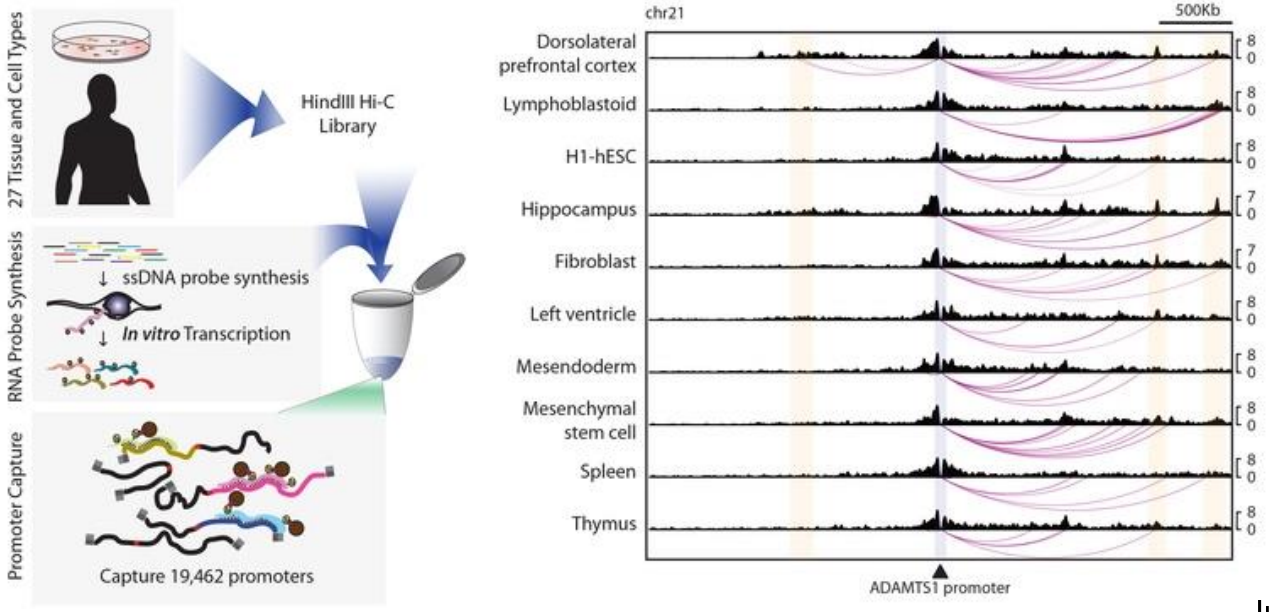


Fulco et al 2019 *Nat Genet*
Nasser et al 2021 *Nature*

# Promoter-capture Hi-C to link elements to genes



Javierre et al 2016 Cell



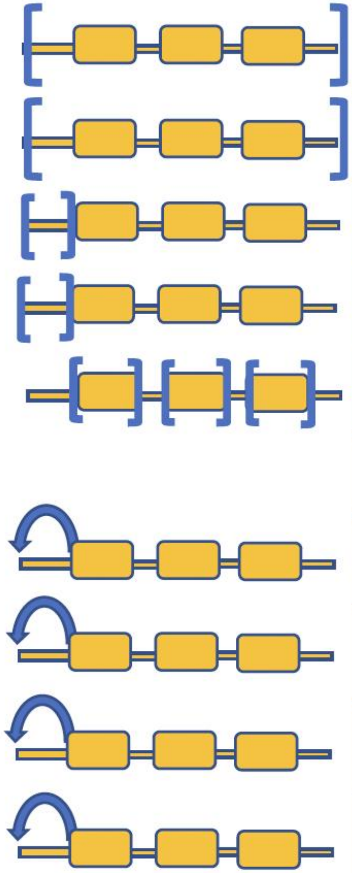
Jung et al 2020 *Nat Genet*

# Broadening the scope of approaches to link variants to genes

Naive S2G

Expression S2G

Hi-C S2G

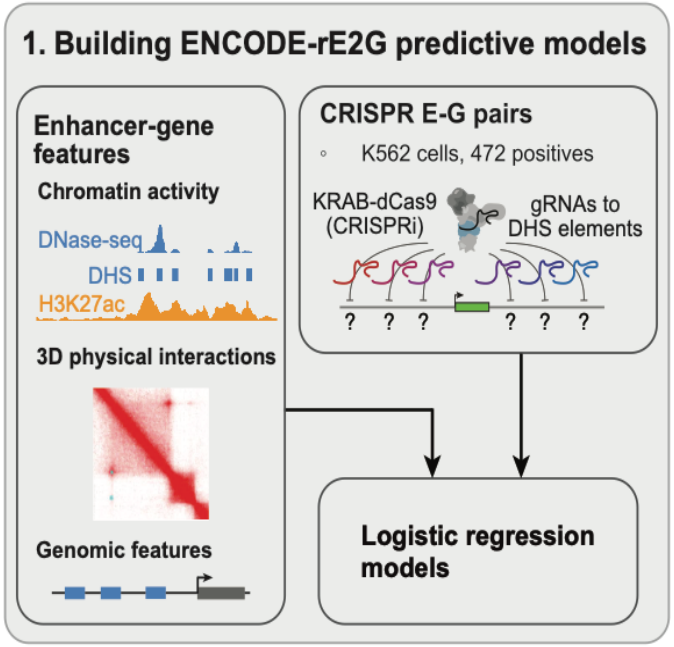


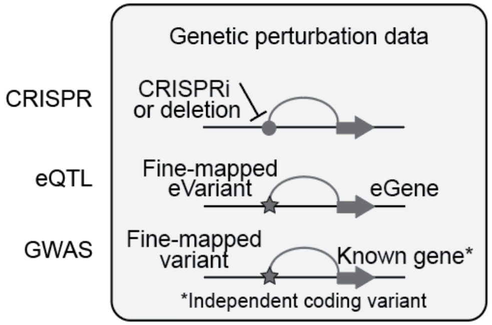
| S2G strategies | Description |
|---|---|
| 5kb | SNPs in 5kb window around gene |
| 100kb | SNPs in 100 kb window around gene |
| Promoter | SNPs in promoter region of the gene |
| TSS | SNPs in  and around Transcription start sites |
| Coding | SNPs in coding regions of the gene |
| eQTL | Max. post. 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] |

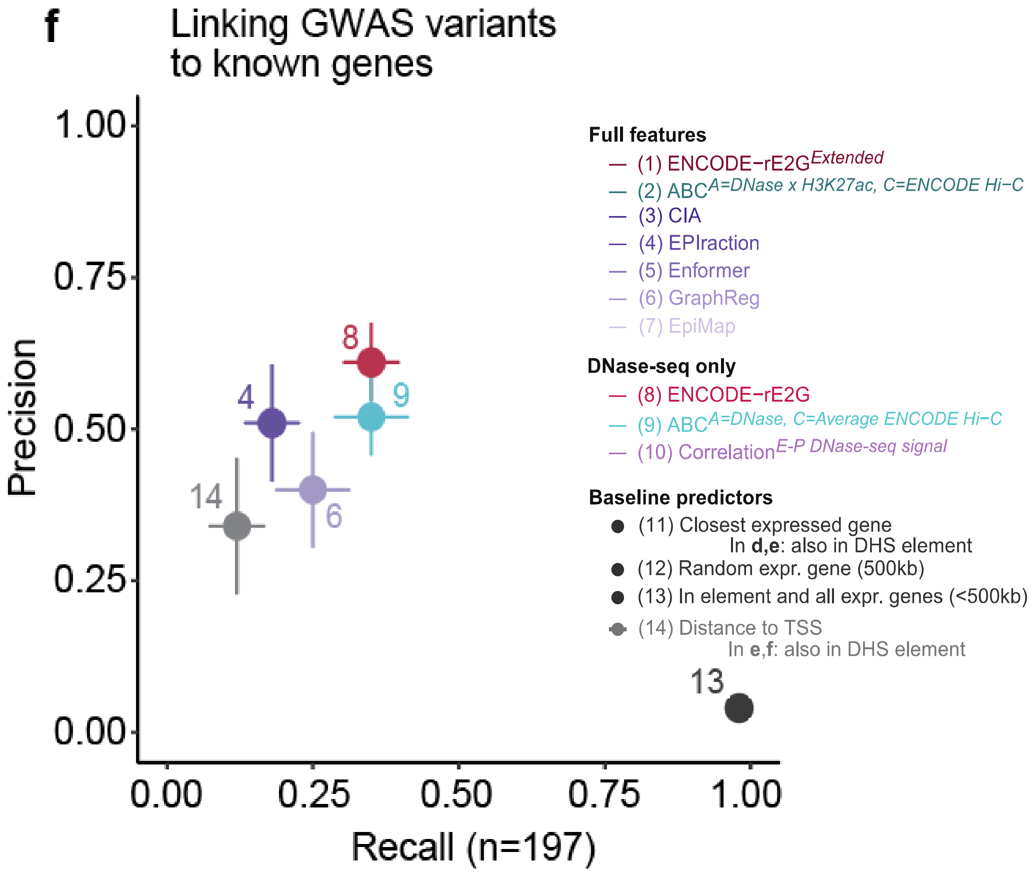
[1]Hormozdiari et al 2018 NG, [2]Aguet et al 2019 bioRxiv, [3]Yoshida et al 2019 Cell, [4]Liu et al 2017 Gen. Biol. , [5]Ernst et al 2011 Nat. meth., [6]Kundaje et al 2015 Nature , [7]Javierre et al 2016 Cell,, [8]Fulco et al 2019 Nat.Genet

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

# Benchmarking different element-gene linking approaches

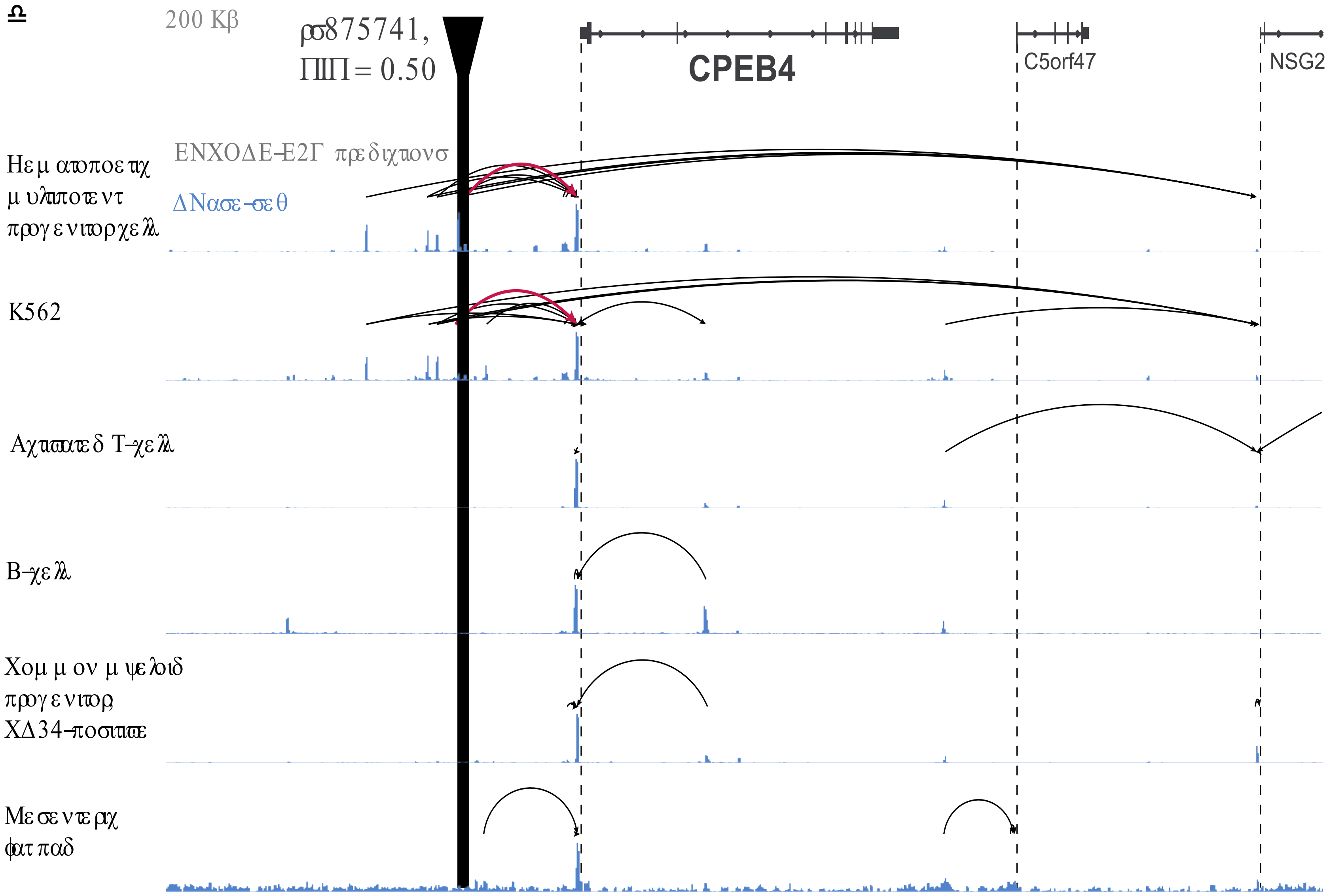






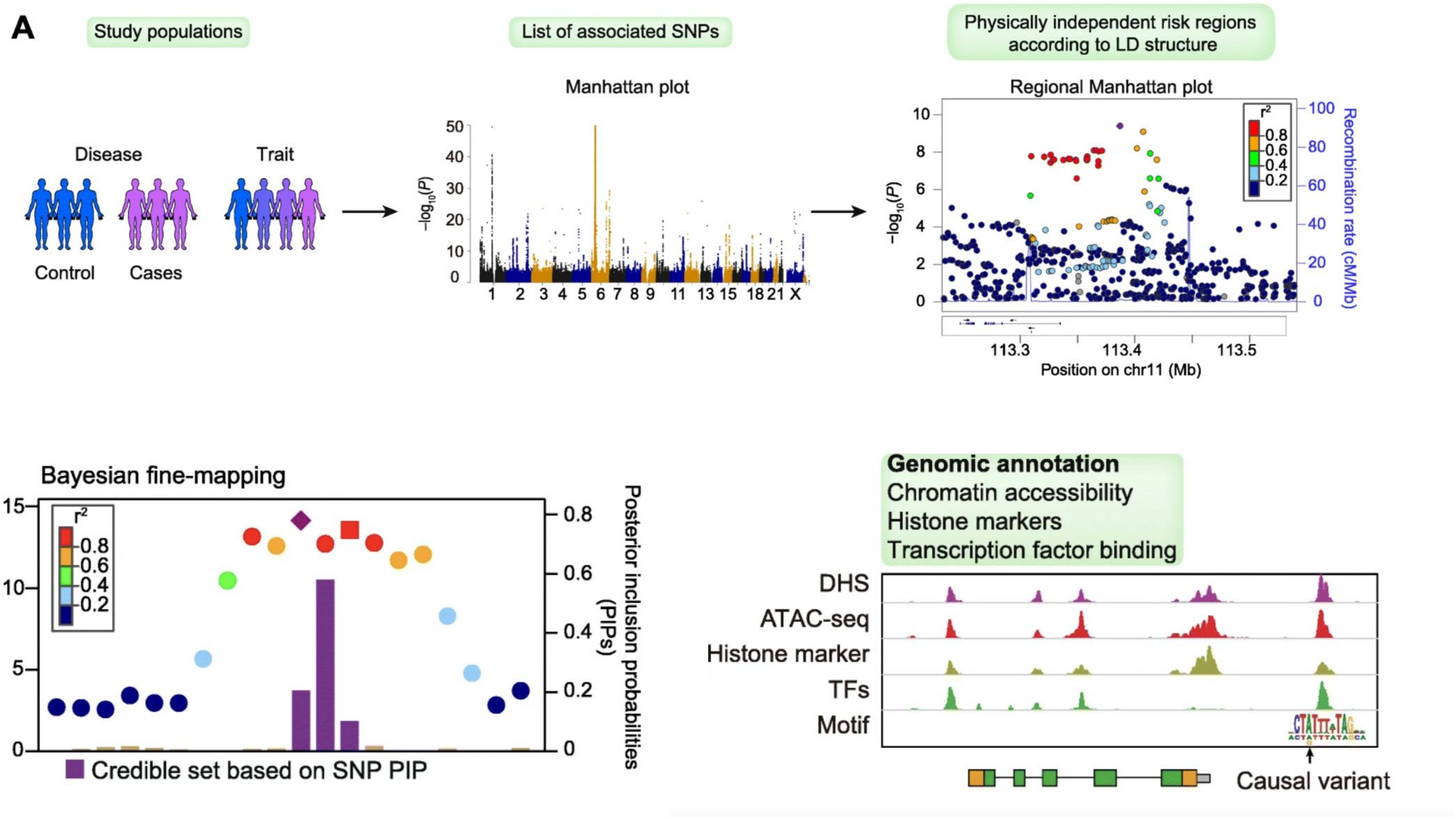
Gschwind*..,Dey*..Engreitz et al 2023 bioRxiv, in rev, *Nature*

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



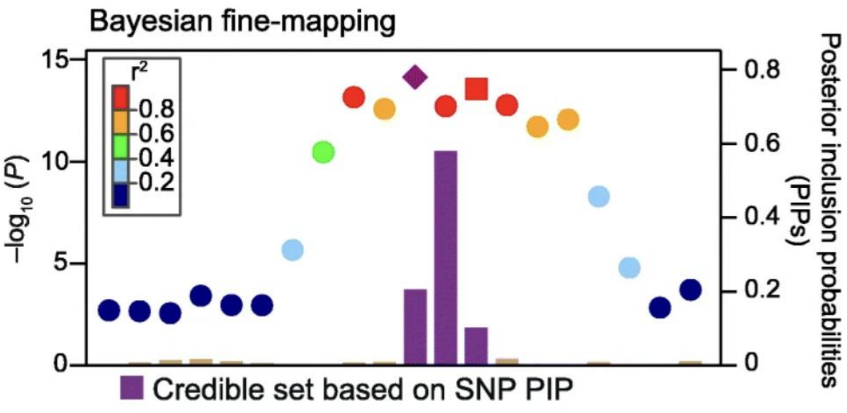
# Artificial perturbation screens for human molecular phenotypes (CRISPR)

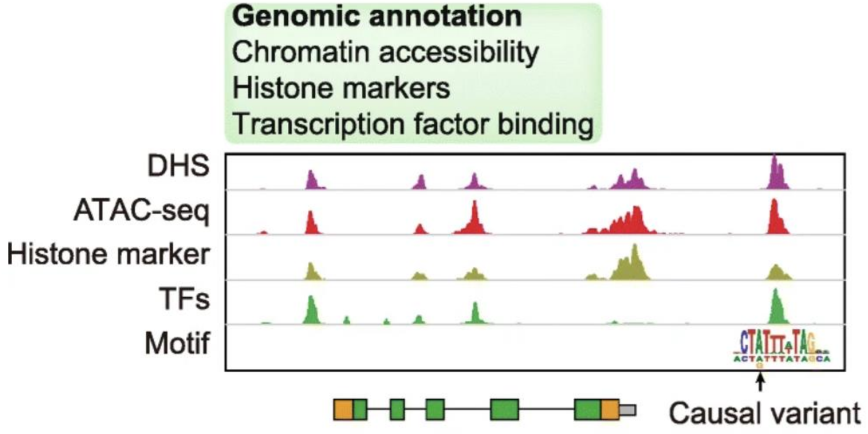
# Functional characterization targeting GWAS risk variants



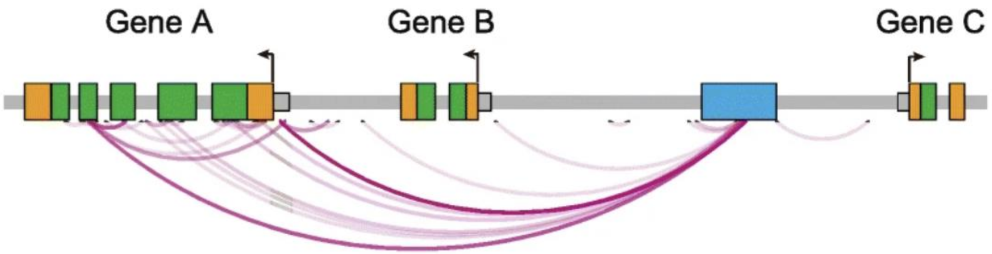
Rao et al 2021, *Genome Medicine*

# Functional characterization targeting GWAS risk variants

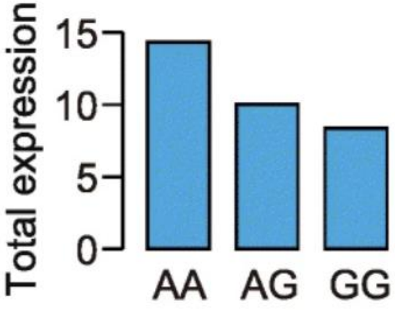


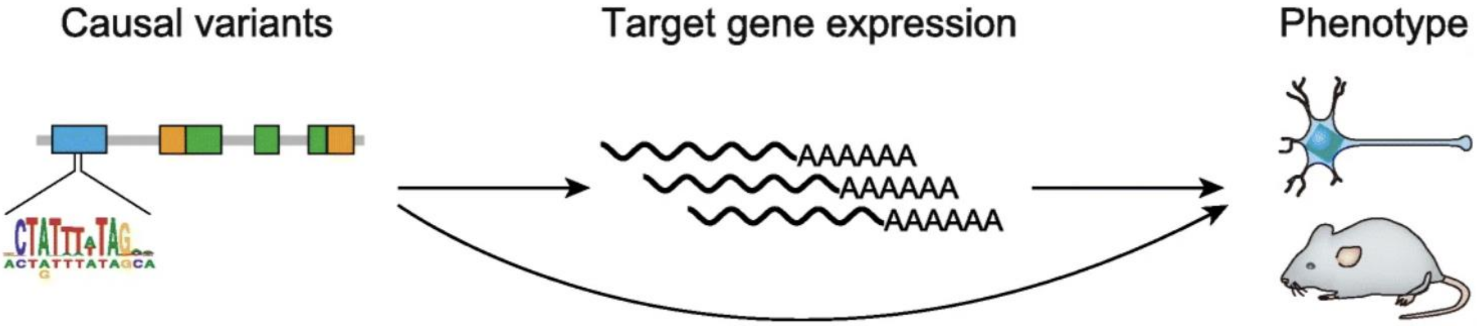


Linking to causal genes (enhancer-gene)



Expression QTL calling



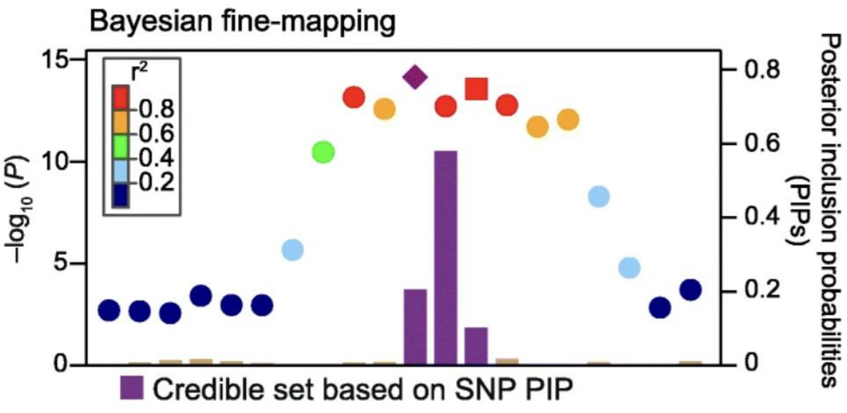


Causal variants

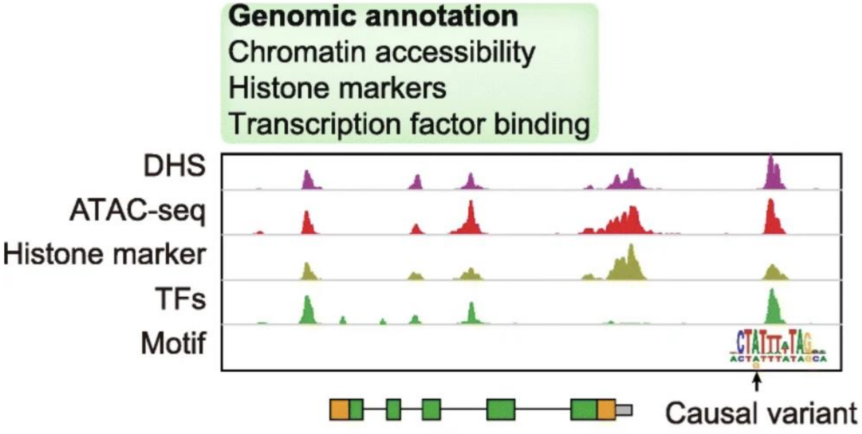
Target gene expression

Phenotype

# Functional characterization targeting GWAS risk variants

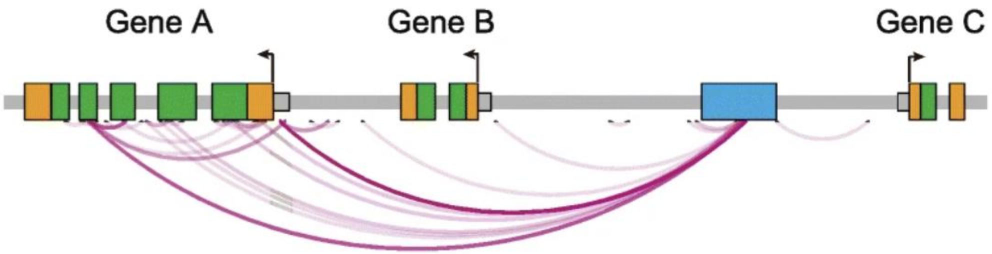

Bayesian fine-mapping

■ Credible set based on SNP PIP

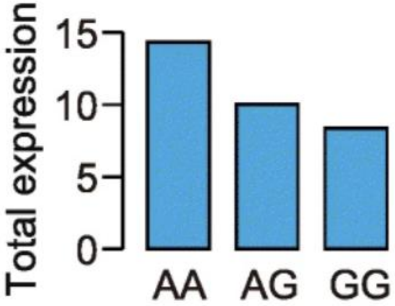

**Genomic annotation**
Chromatin accessibility
Histone markers
Transcription factor binding

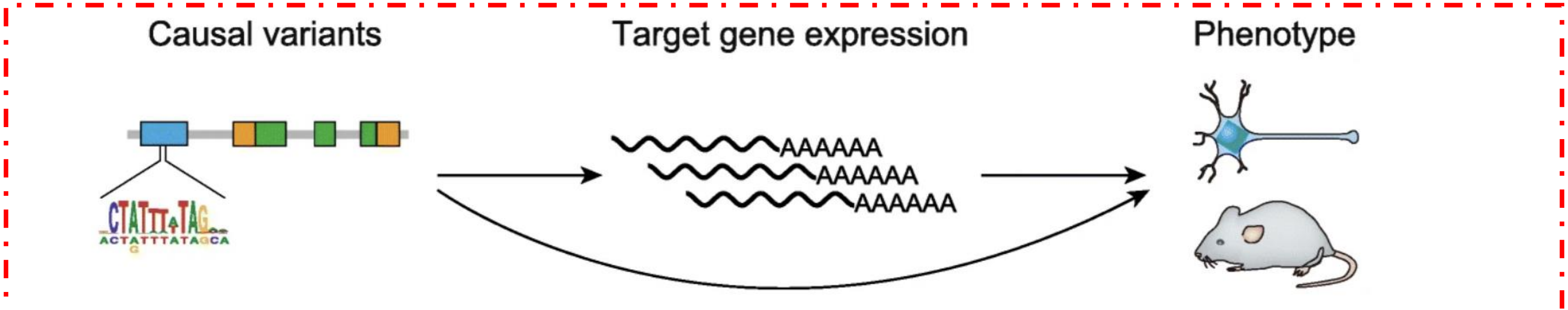
Causal variant

Linking to causal genes (enhancer-gene)



Expression QTL calling

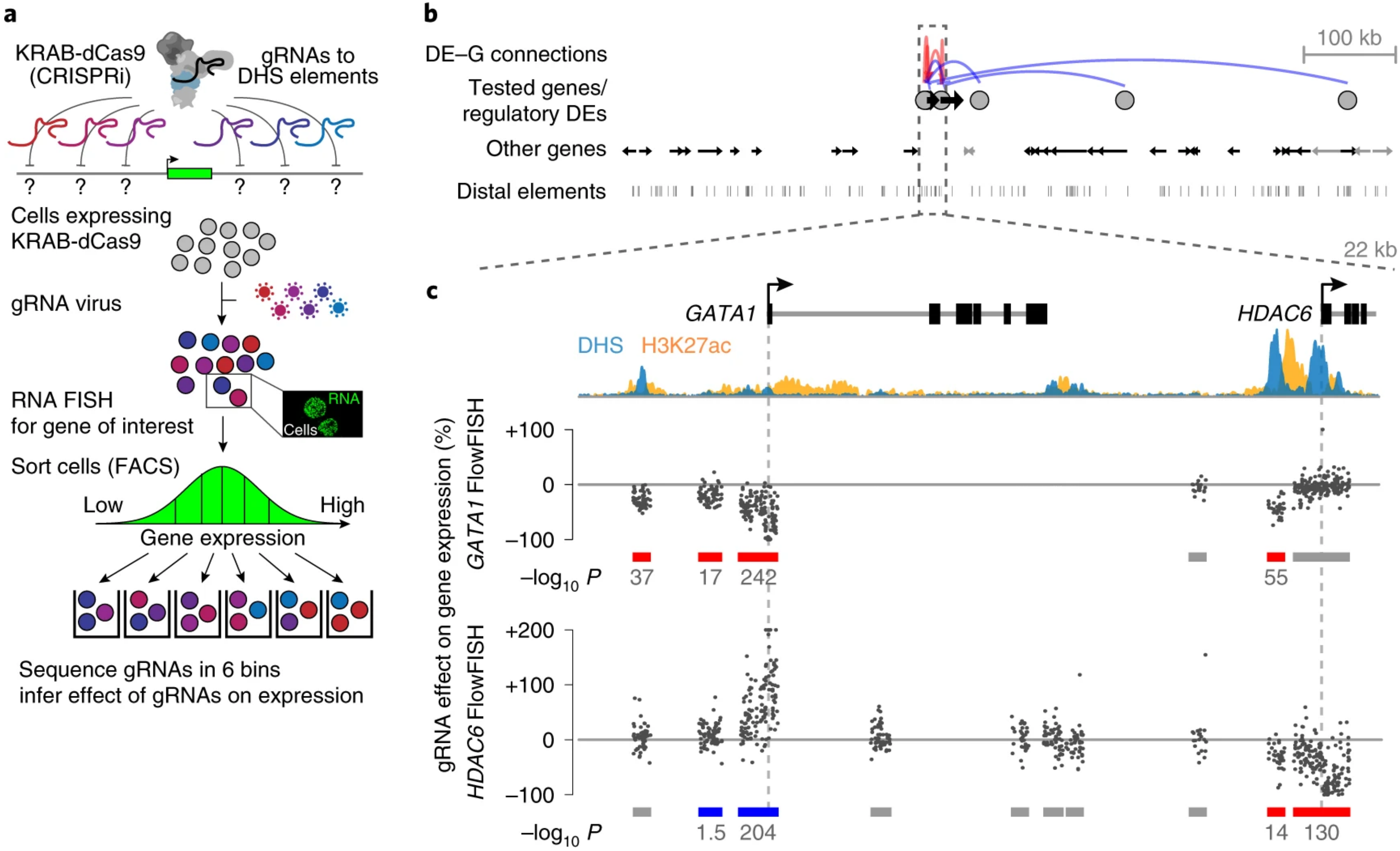



Causal variants

Target gene expression

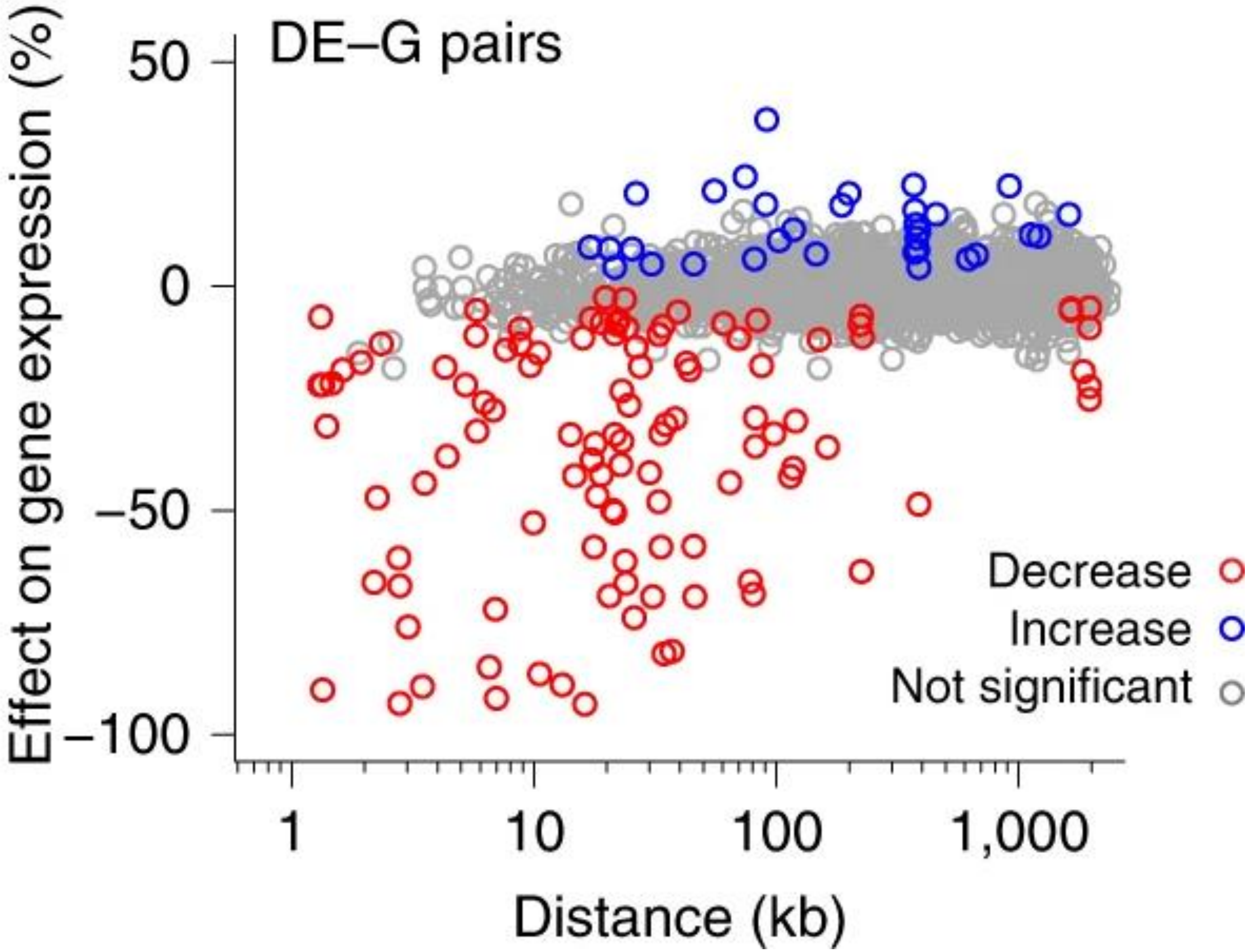
Phenotype

# CRISPRi perturbation screen in K562 mimic-ing cis eQTLs



Fulco et al 2019, *Nat Genet*

# CRISPRi perturbation screen in K562 mimic-ing cis eQTLs



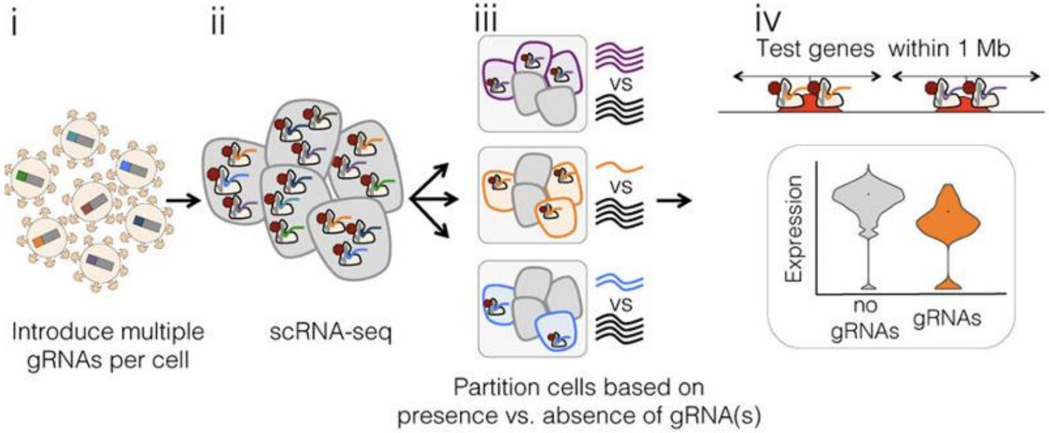
Fulco et al 2019, *Nat Genet*

# Large scale genome-wide enhancer perturbation screen to mimic cis and trans eQTLs

CRISPRi Perturb-seq (TSS-targeted or enhancer–targeted): dCas9-KRAB, can assess global changes in transcriptomic profile owing to one or sets of perturbations.

By introducing gRNAs at a high MOI (~30), each individual cell acquires a unique combination of perturbations, which markedly increases statistical power.

Incorporating in low MOI (<= 1~2) however enables more accurate understanding of a single perturbation effect

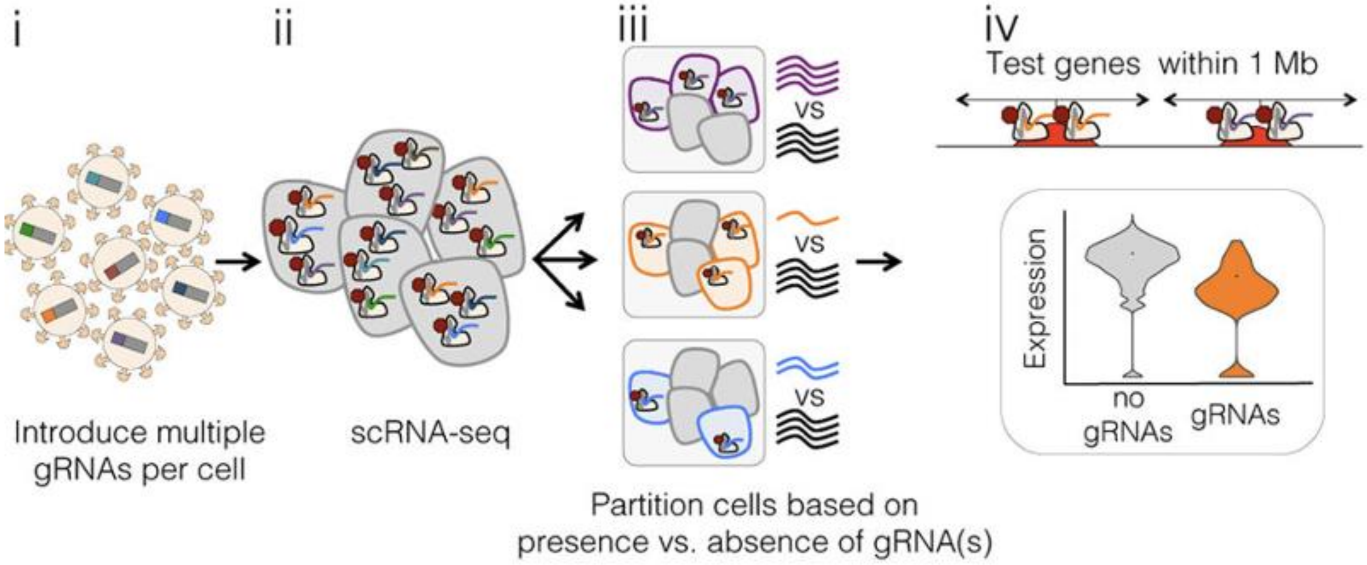


i
ii
iii
iv
Test genes within 1 Mb
vs
vs
vs
Expression
no gRNAs
gRNAs
Introduce multiple gRNAs per cell
scRNA-seq
Partition cells based on presence vs. absence of gRNA(s)

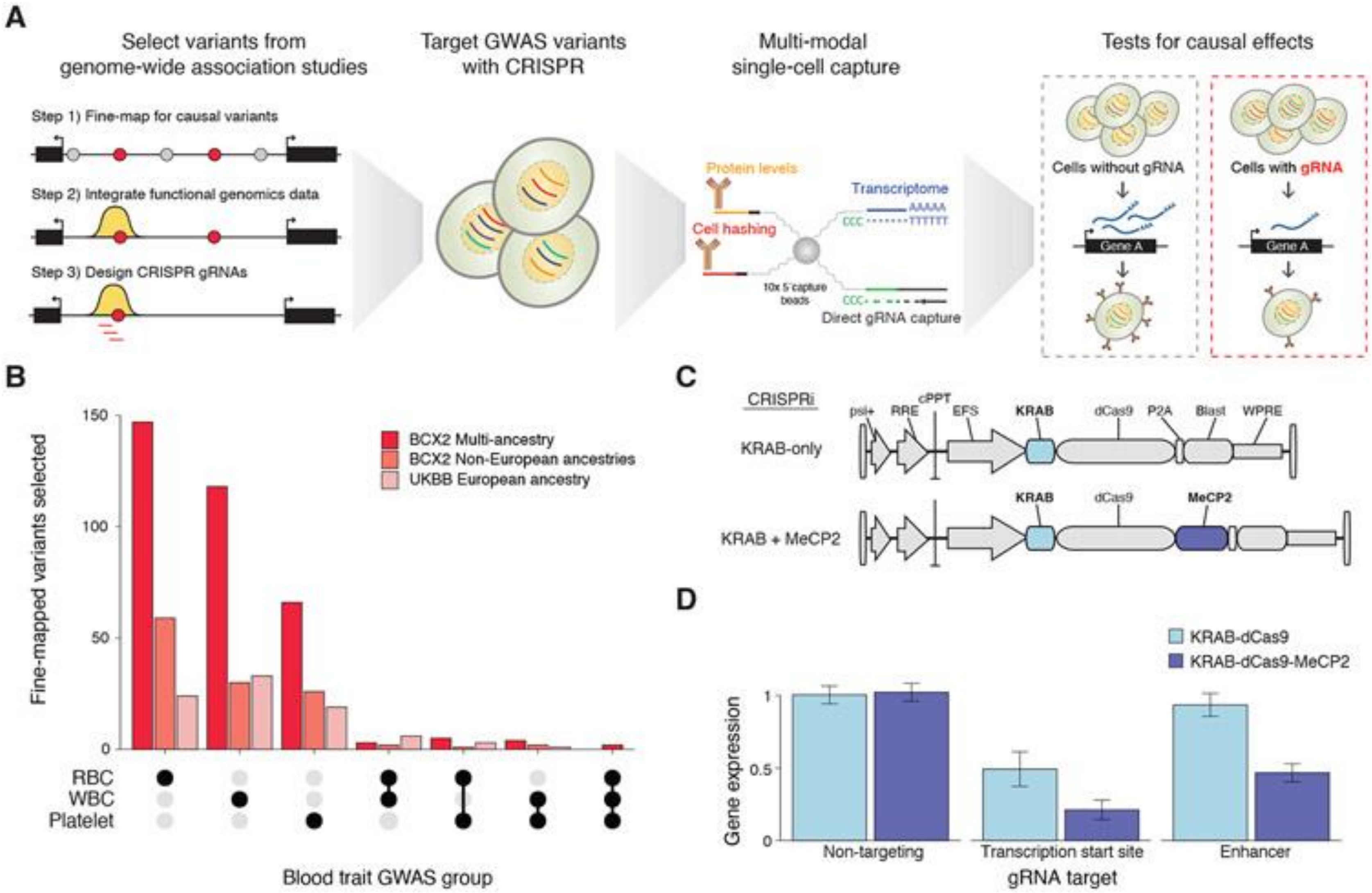
# Large scale genome-wide enhancer perturbation screen to mimic cis and trans eQTLs

Map *cis* and *trans* effects by comparing gene expression in the subset of cells that contain a given gRNA to those that lack that guide (similar to eQTL) (crisprQTL mapping).

Unlike eQTL studies, the resolution of our screen is not constrained by linkage disequilibrium, nor is it limited to studying sites in which common genetic variants happen to exist.

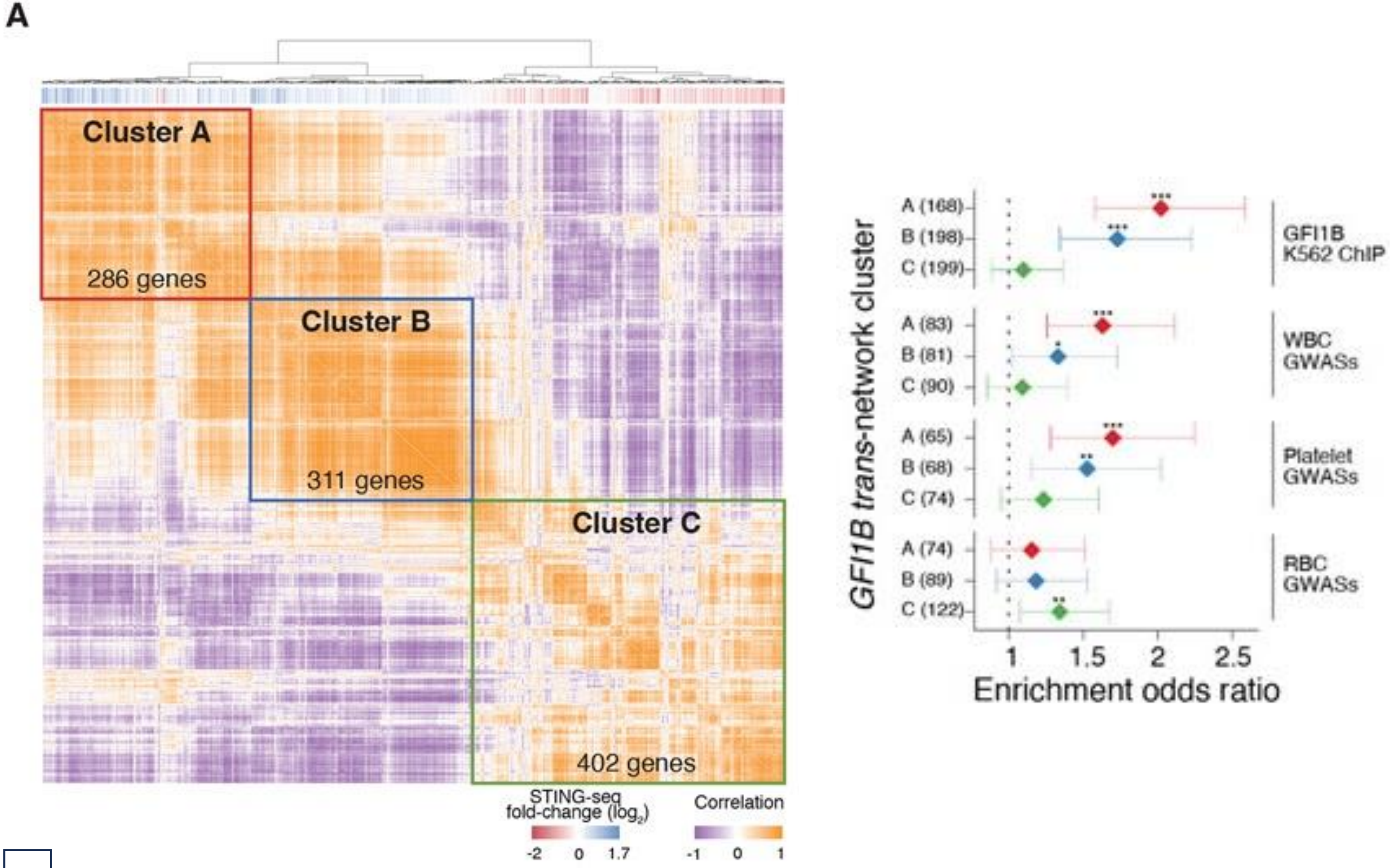

i
ii
iii
iv
Test genes within 1 Mb
vs
vs
vs
Expression
no gRNAs
gRNAs
Introduce multiple gRNAs per cell
scRNA-seq
Partition cells based on presence vs. absence of gRNA(s)

# STING-seq: CRISPRi and CRISPR base-editing efforts targeting variants fine-mapped from immune traits



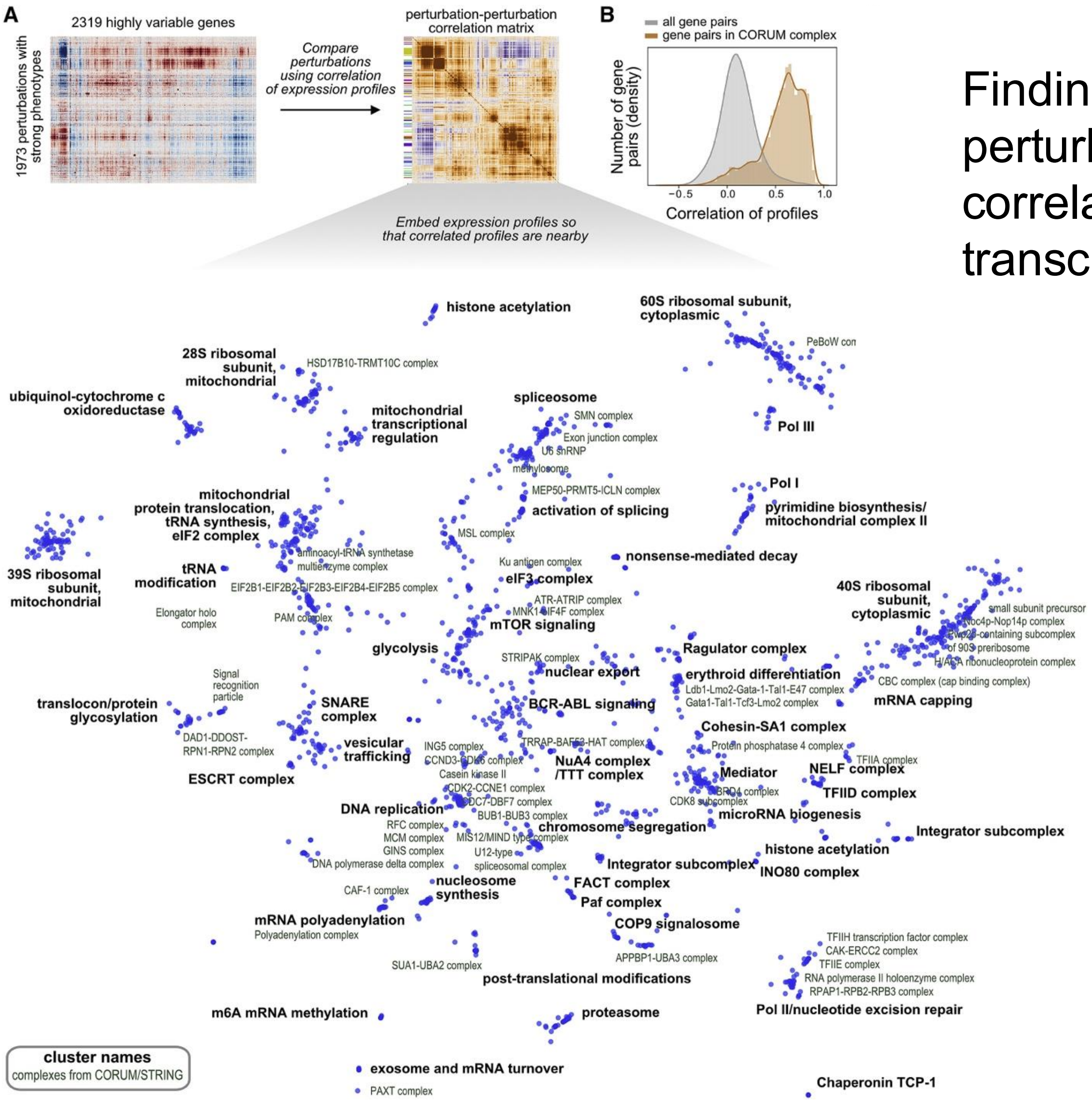
Sanjana lab NYGC, Morris et al 2023 *Science*

# STING-seq: CRISPRi trans-effect hubs similar to trans-eQTL programs of genes



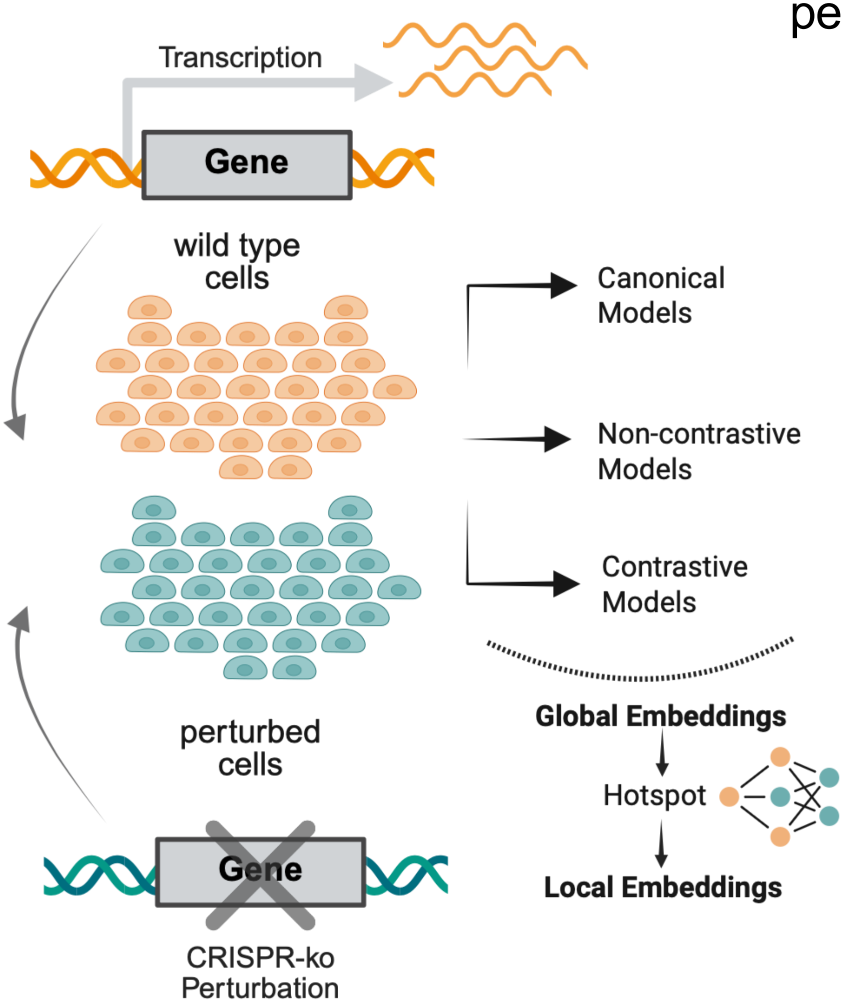
Sanjana lab NYGC, Morris et al 2023 *Science*

# Defining programs of genes underlying CRISPR perturbations

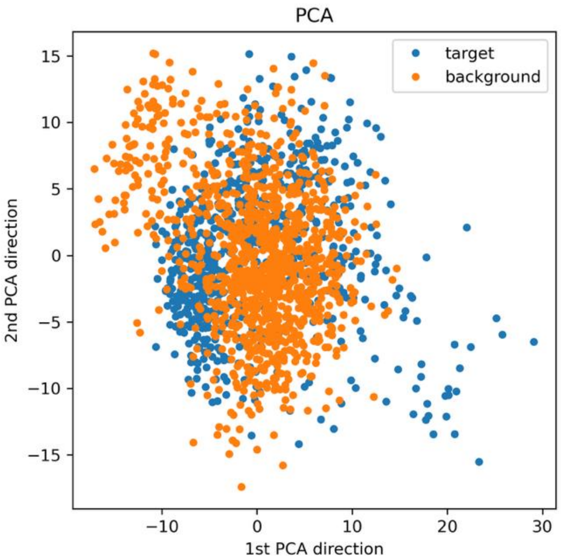


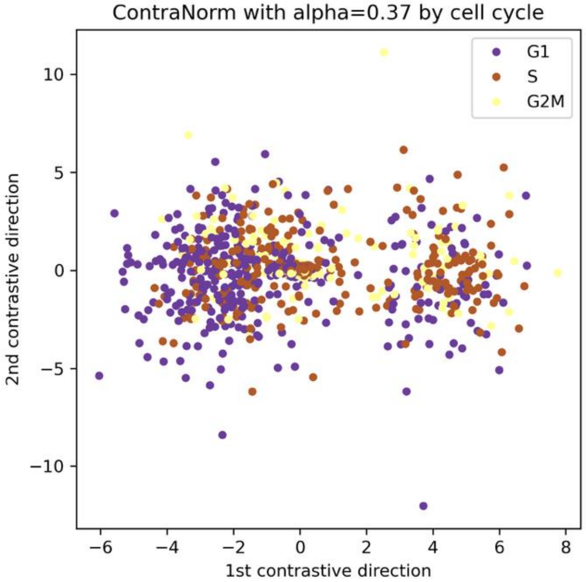
Finding hubs of perturbations with correlated transcriptomic effects

# Contrastive embedding approaches often find interesting structure among genes modulated by a perturbation



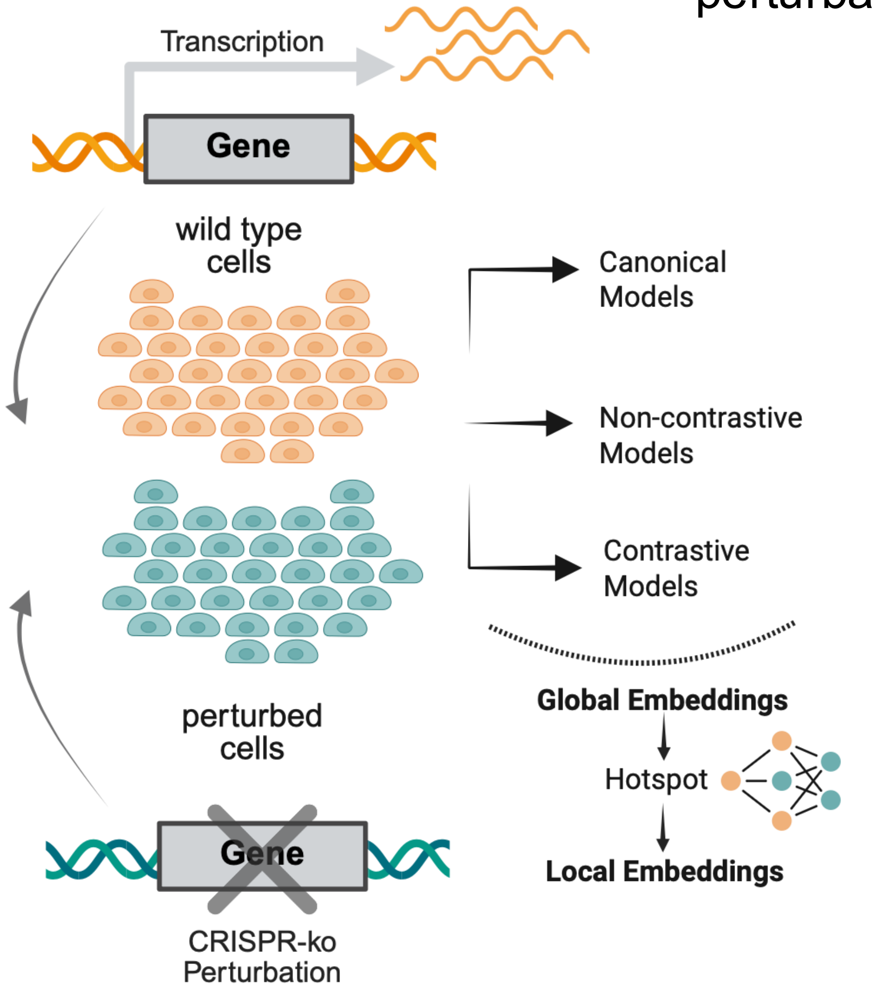
PDX1 perturbation

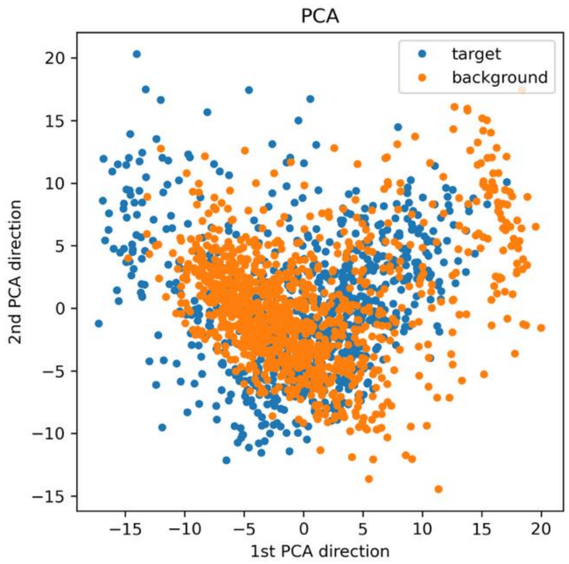


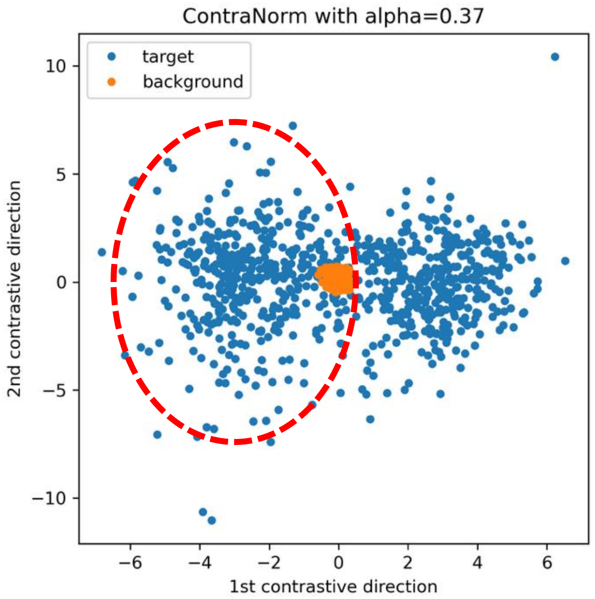


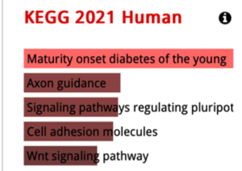
# Contrastive embedding approaches often find interesting structure among genes modulated by a perturbation

GATA6het perturbation



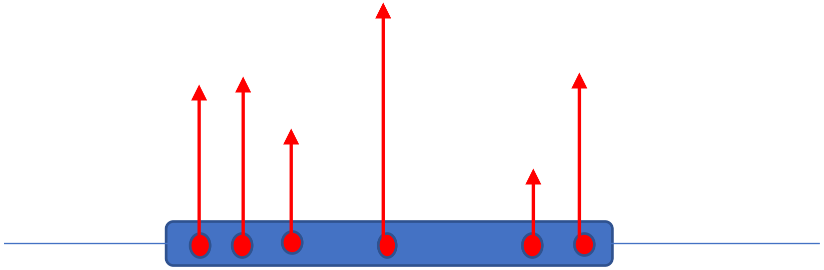






# Linking Gene Programs to Disease (G2D) from perturbation screens

# Prioritizing genes for a complex disease (MAGMA and PoPS)



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

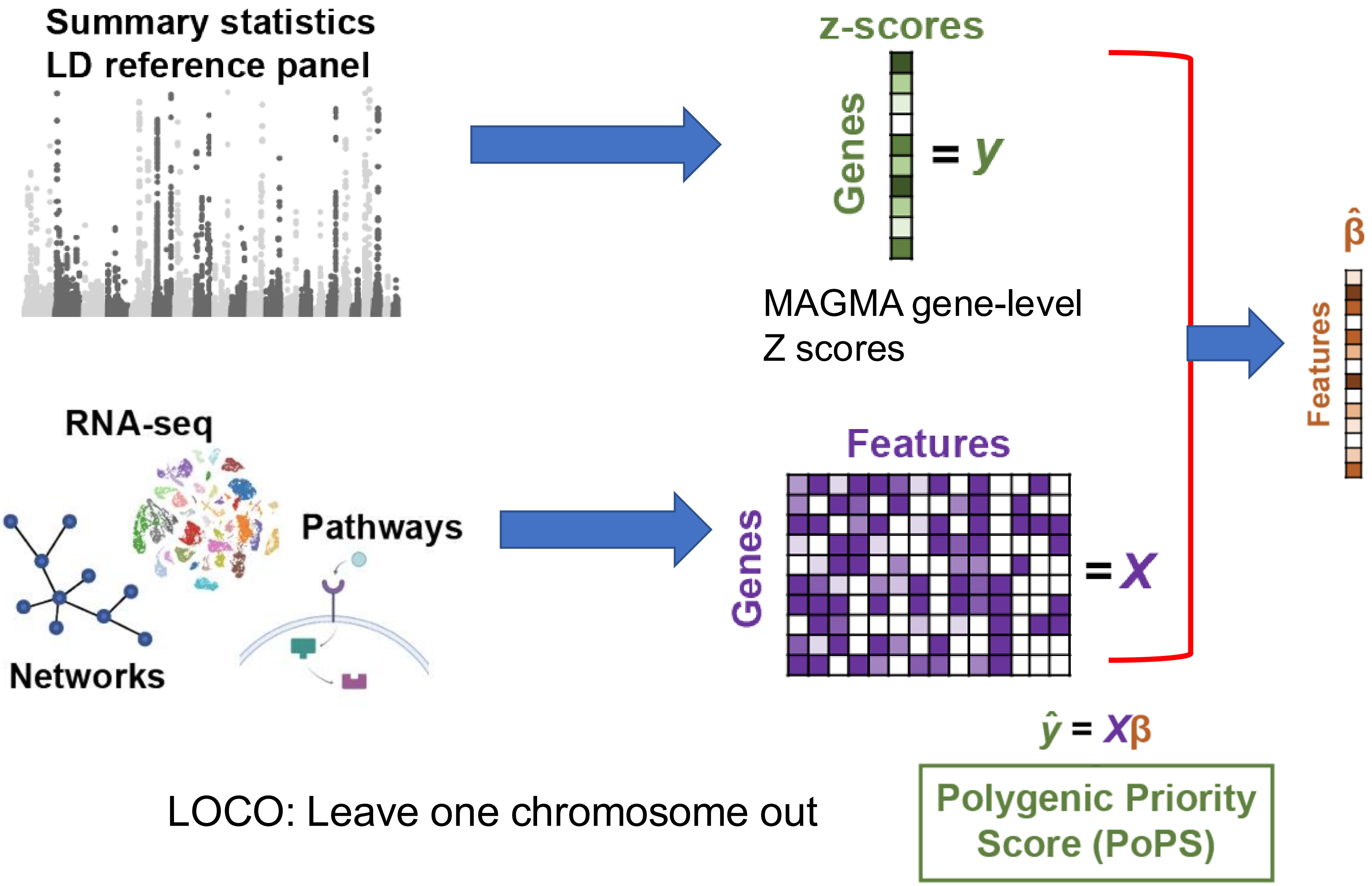
(a) The mean of the $\chi^2$ statistic for the SNPs in a gene,

(b) The top $\chi^2$ statistic among the SNPs in a gene.

For the mean $\chi^2$ 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*



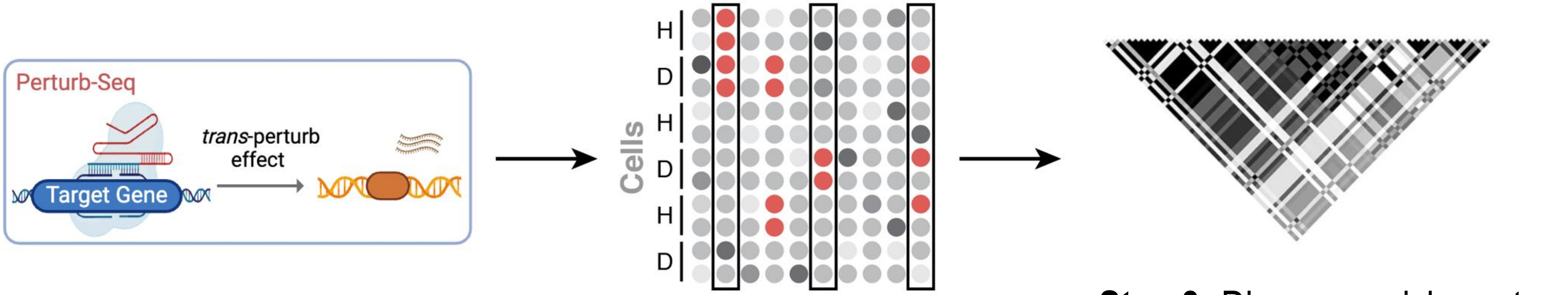
MAGMA gene-level Z scores

$\hat{y} = X\beta$

Polygenic Priority Score (PoPS)

LOCO: Leave one chromosome out

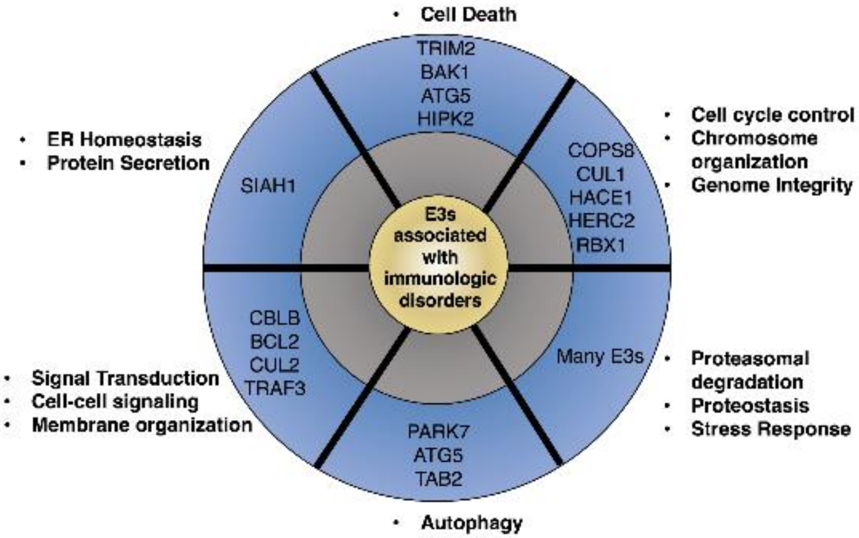
# Disease information in Perturb-seq co-regulated gene programs



**Step 1:** Gene programs

**Step 3:** Disease enrichment

Enrichment of PoPS/MAGMA or any score in your perturbation program compared to similar-sized program drawn from genes affected by >=1 perturbation

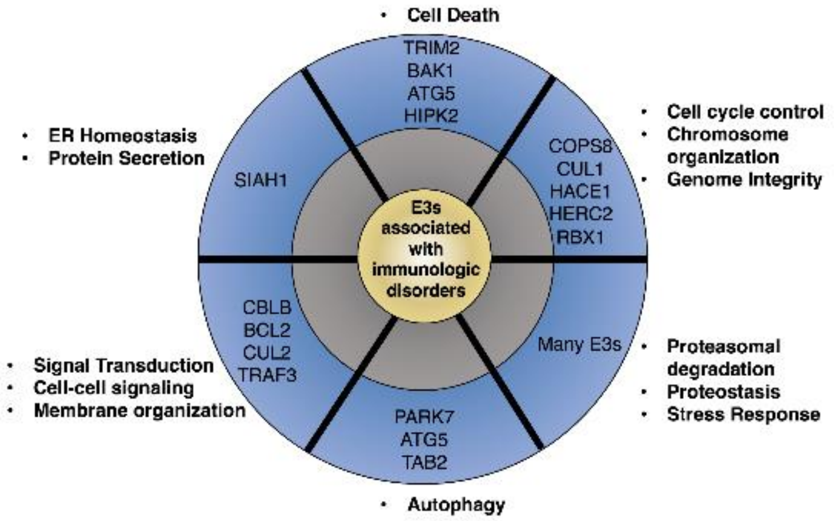


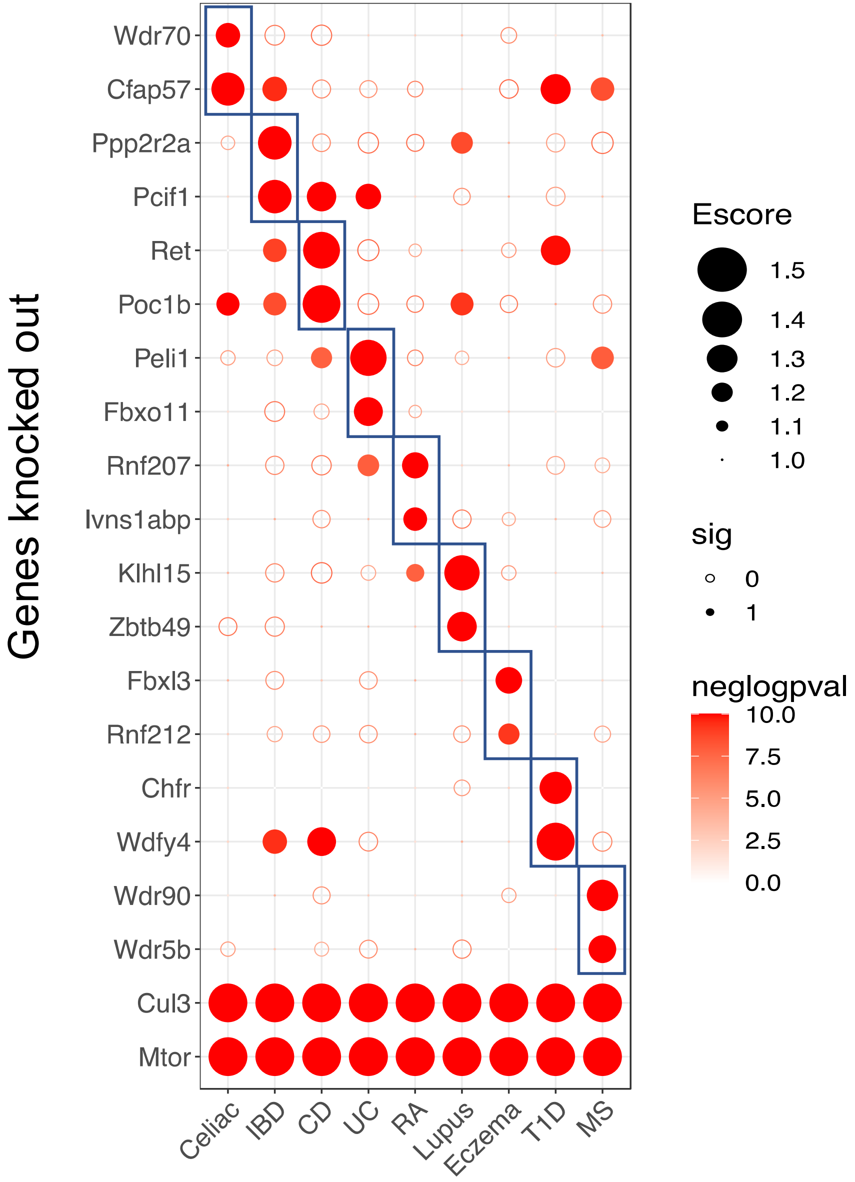
Large-scale Perturb-seq assay targeting 1031 E3ligase genes in bone marrow dendritic cells

Geiger-Schuller, Eraslan et al bioRxiv 2023, in rev *Cell*

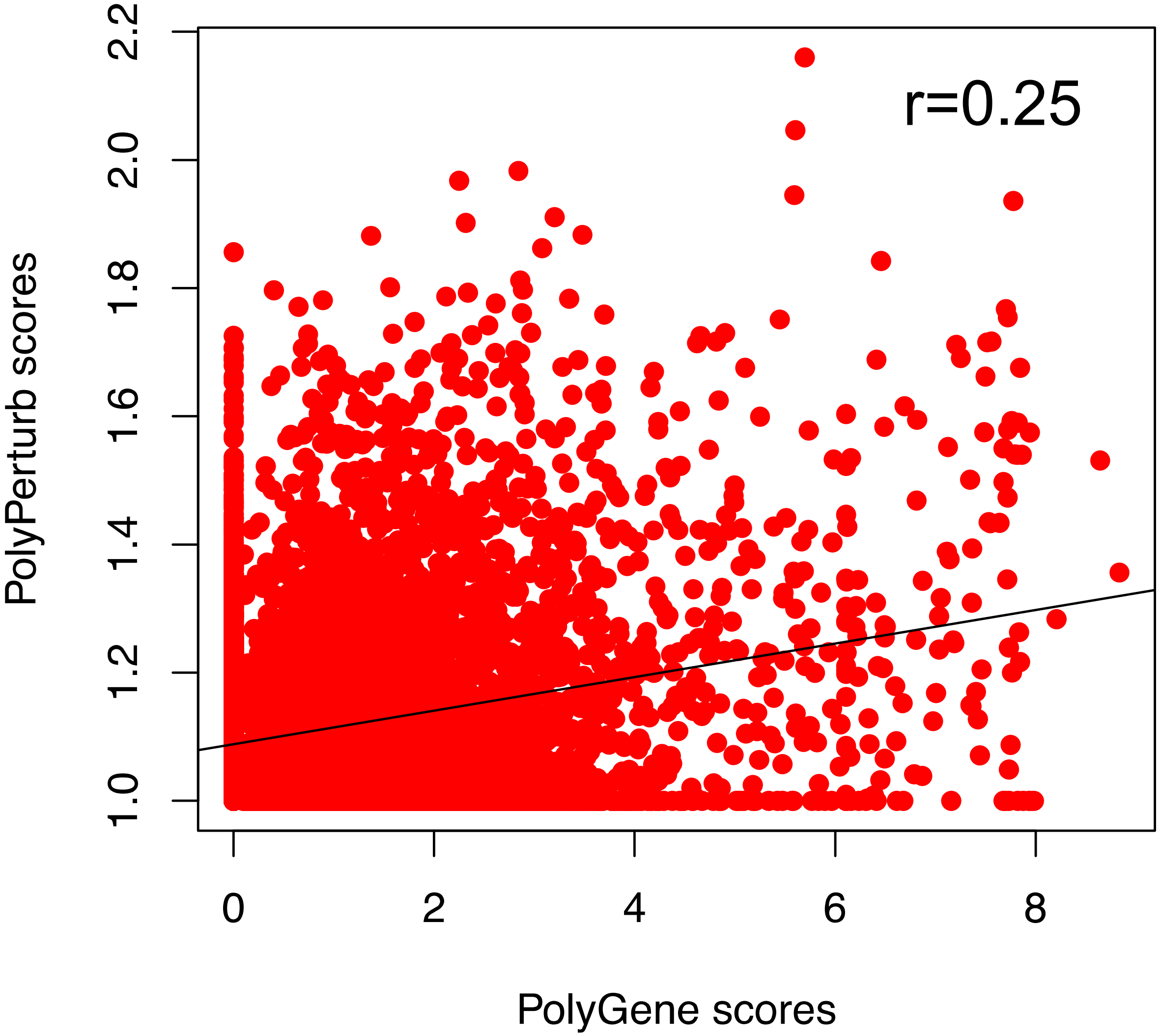
# GeneBoost approach to score perturbation programs for disease

1,030 E3 ligase genes and interacting partners





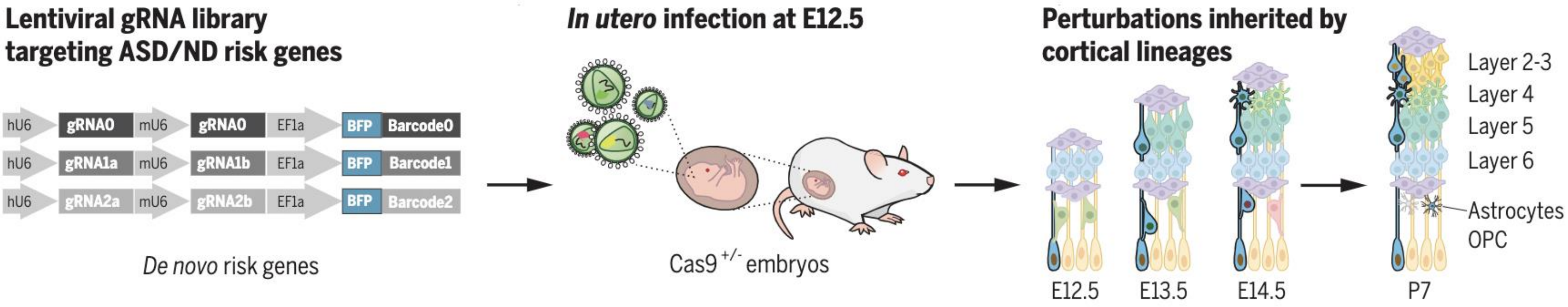
# PoPs gene-level scores of knockouts and PoPs enrichment of their perturbation profiles are moderately correlated

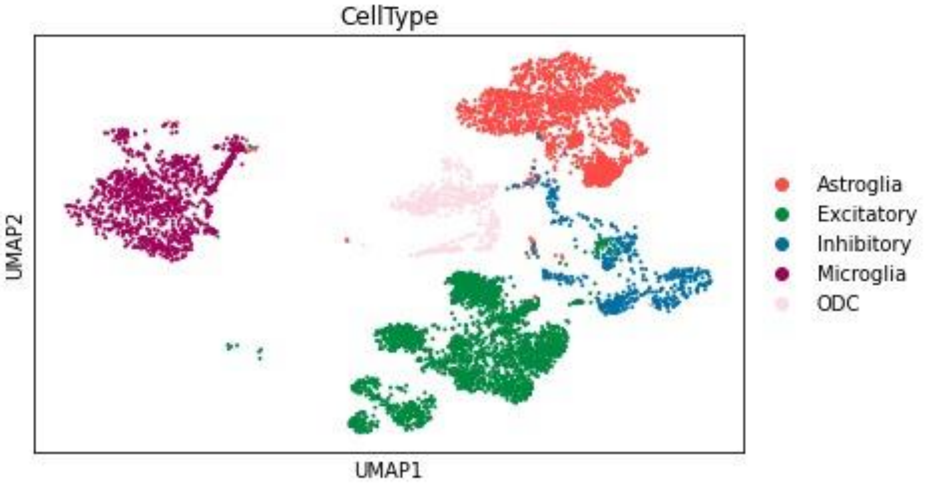


Each point is a (KO gene, immune disease) pair in E3ligase Perturb-seq experiment

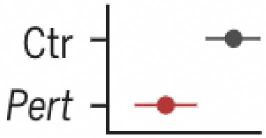
We consider 1,030 genes and 9 immune related traits.

# In-vivo Perturbation programs across multiple cell types (Jin et al Science 2020)





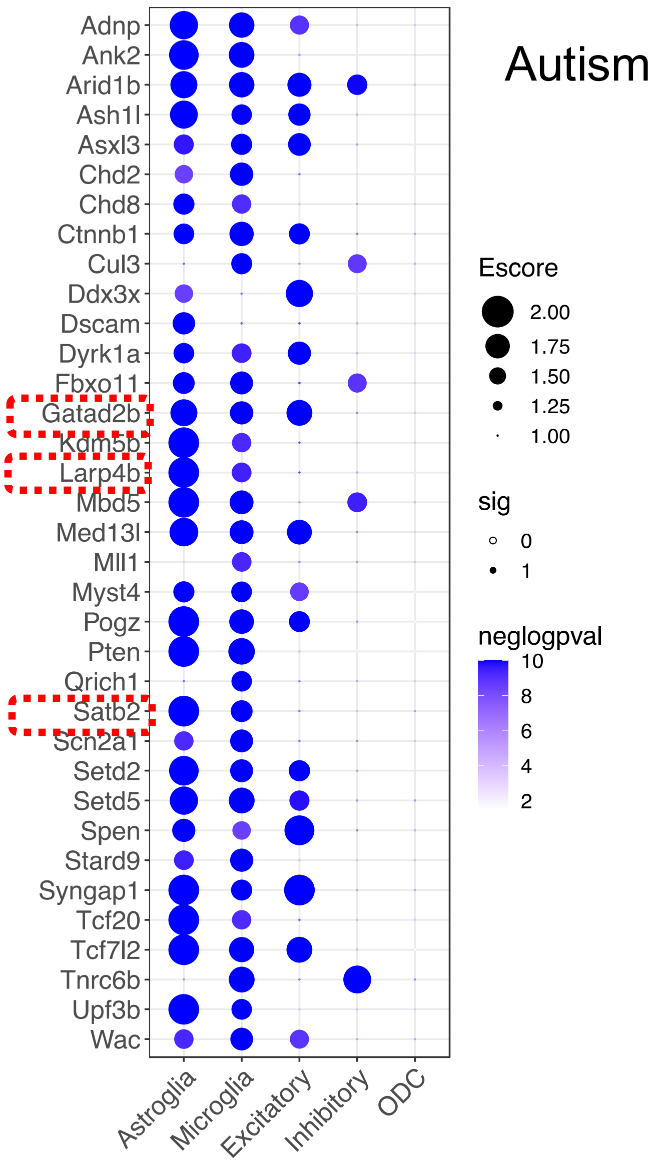
35 de-novo autism risk genes targeted

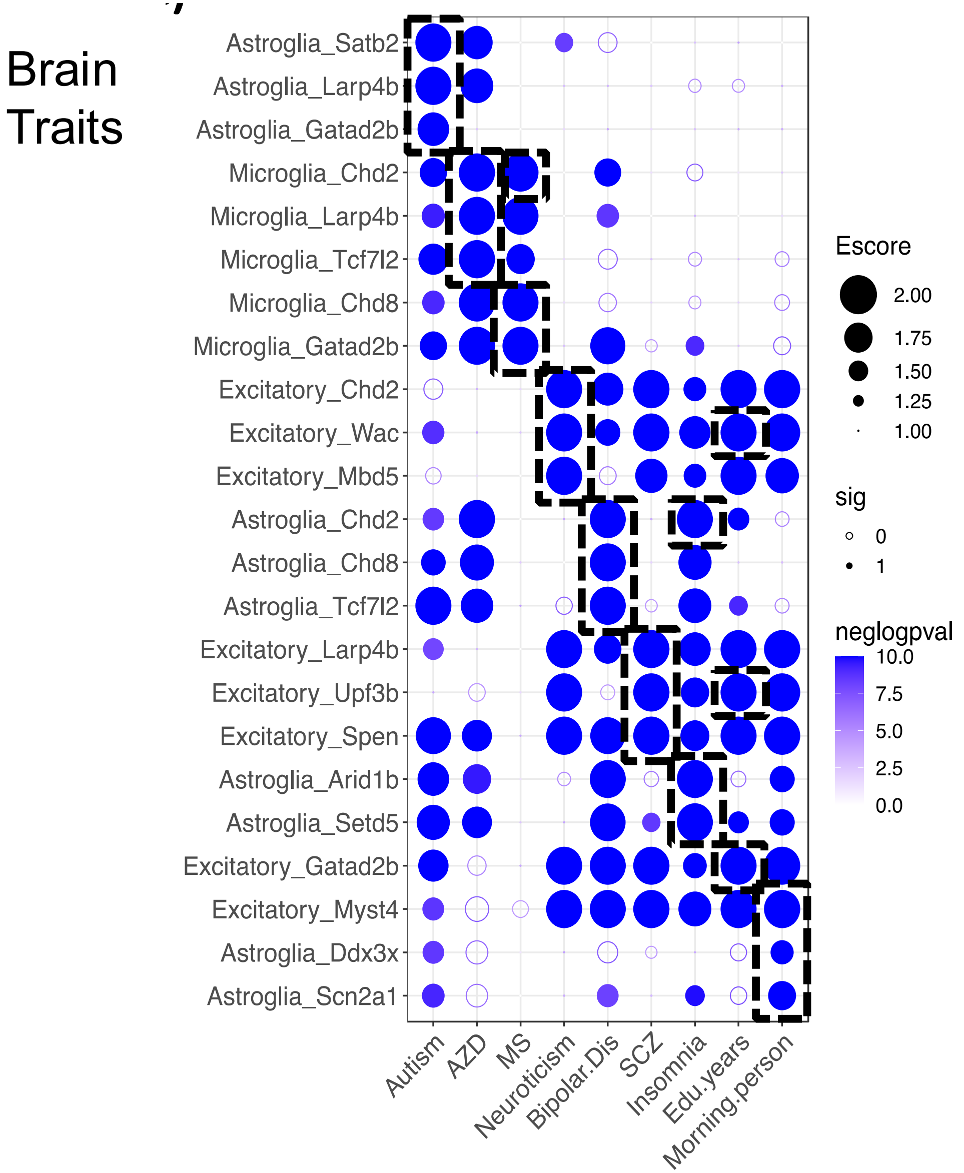


Perturbation program observed for each guide in 5 major brain cell types for a total Of 175 perturbation programs.

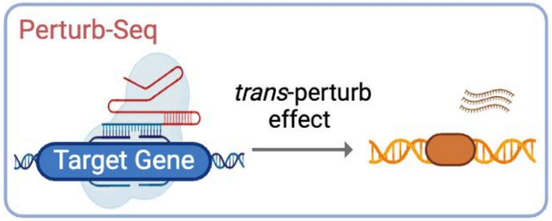
Jin et al Science 2020

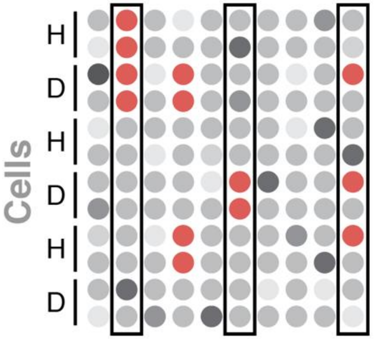
# Some Astroglia perturbation programs are specifically disease informative for autism compared to other brain related diseases



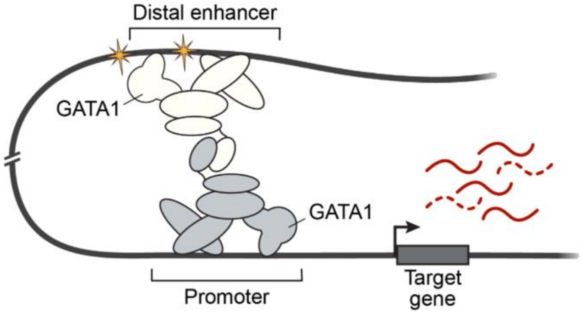


# sc-linker heritability analysis of Perturb-seq co-regulated gene programs





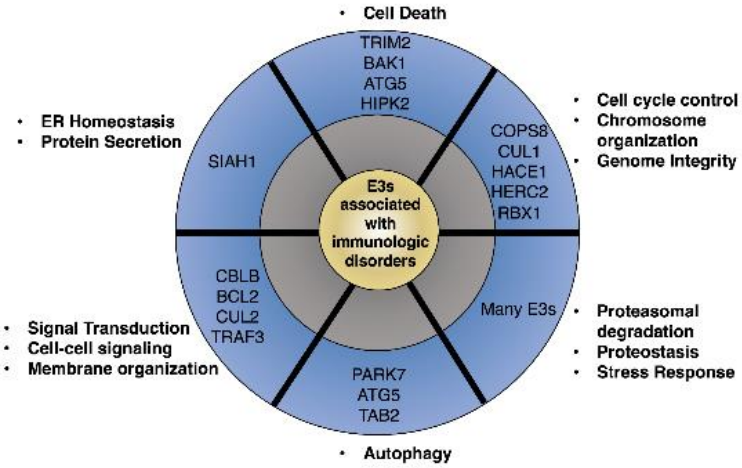
**Step 1:** Gene programs



**Step 2:** SNP-gene maps to generate SNP annotation

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

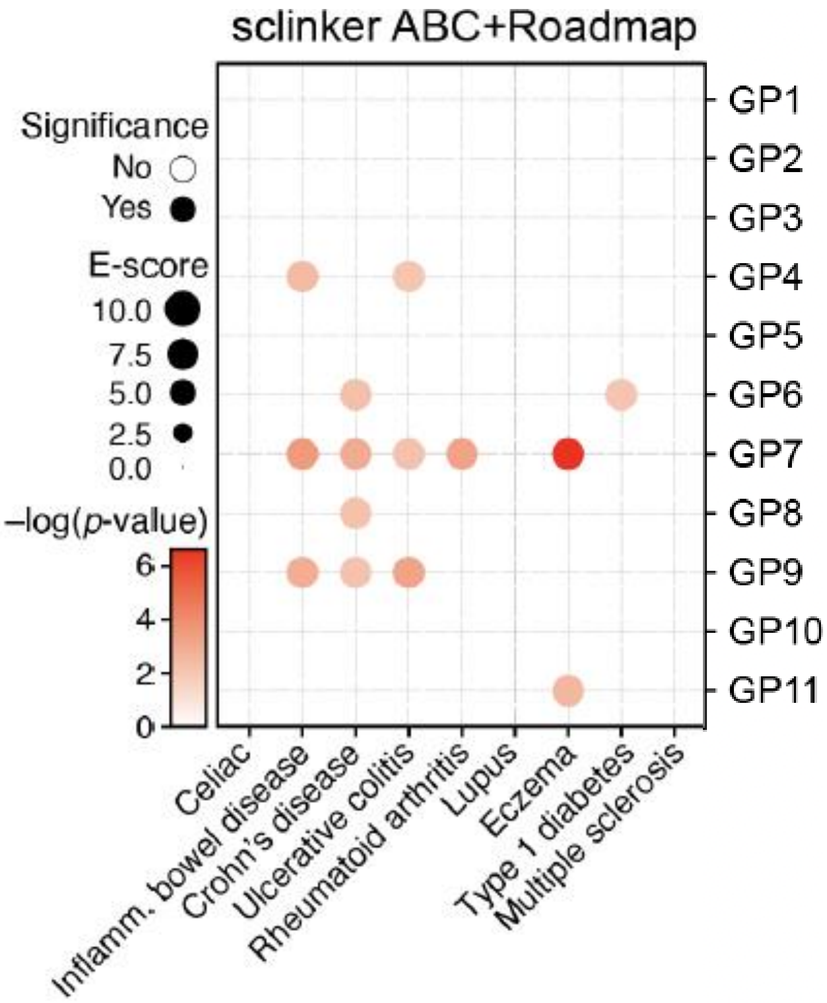
**Step 3:** Disease heritability enrichment



Large-scale Perturb-seq assay targeting 1031 E3ligase genes in bone marrow dendritic cells

Geiger-Schuller, Eraslan et al bioRxiv 2023, in rev *Cell*

# We observe specific immune disease heritability enrichment using sc-linker in various Perturb-seq programs



**Program GP 1: Response to oxidative stress**
*Itgam, Itgb2, Acod1, Cd36, Mmp8, Thbs1, Srxn1, Prdx1, Txnrd1, Tpm1, Cat, Gsr, Hmox1, Prdx6, Csf1r, Cxcl3, Gsn* (may be a type 'Gsr'), *Clec5a, Msr1, Bst1*

**Program GP 2: Response to ER stress**
*Selenos, Surf4, Sec11c/22b/61b/61g, Pdia3/4/6, Herpud1, Hsp90b1*

**Program GP 3: Pyruvate metabolism**
*Tpi1, Pgam1, Eno1, Hk2, Hk1, Pfkl, Ldha, Pkm, Bsg, Pgk1,Aldoc, Aldoa, Gapdh, Slc16a3*

**Program GP 4: Motility and cell maintenance**
*C3ar1, Ccl2/7, Cdh1, Map1lc3b, Pdlim7, Plxnb2, Spata13, Swap70, Vim, Snrpf, Snrpd2, Nop58, Eif3e/f/i/k, Trem1/2, Hnrnpa1*

**Program GP 5: Protein homeostasis and phagocytosis**
*Hsp90ab1, Hspa8, Ubb, Nedd8, Ube2m, Vcp, Psma4/5/6/7,Actb1/g1, Actg1, Arpc1b, Coroa1, Tubb1a/1b/5, Ppia, Tyrobp, Atp5/Cox/Uqcr family genes, Erp29, Reep5, Ssr4, Krtcap2*

**Program GP 6: Ribosome / translation**
*Rpl3, Rps26, Rps20, many other Rpl/Rps genes,Rack1, Npm1, Tpt1, Naca*

**Program GP 7: mDC**
*Nfkb2, Il12b,Cd83, Icosl, Icam1, Jak2, Atf5, Ccl22, Ccl5, Marcks, Nfat5, Stat5a, Nfkbia/z, Rel, Itgal, Ikbke, Cd274*

**Program GP 8: TNF / LPS response**
*Cd33, Cd38, Cxcl1/2, Cybb, Gas7, Gng12, Gpr84, Il1a, Il1b, Nlrp3, Sirpa, Syk, Tlr2, Tnf, Il18*

**Program GP 9: Regulation of autophagy and inflammation**
*Cd84, Ly75, Ccl6, Cd63, Cd68, Ctsa/b/c/d/z, Plk2, Psap, Gpr137b, Mcl1, Cd44, Gpnmd, Mt1/2, Fth1, Il7r, Litaf, Mgll*

**Program GP 10: MHC-I Ag presentation**
*B2m, Tapbp, Grn, Hif1a, H2.D1, H2.K1, H2.T23, Lamp1/2, Irf8, Cst3, Ctsk/l/s, Mdm2*

**Program GP 11: DC2 MHC-II Ag presentation**
*H2.Aa, H2.Ab1, H2.DMa, H2.DMb1, H2.Eb1, Cd74, Irf4,Ccr1/5, Ccl17, Socs2, Dcstamp, Slamf9, Itgax, Mgl2, Axl, Anxa1*

Geiger-Schuller, Eraslan et al bioRxiv 2023, in rev *Cell*

# Assignment Problem

Whole genome CRISPRi Perturb-seq data :

Map fine-mapped GWAS variants for K562-related traits to genes using cS2G method and the nearest TSS distance.

Find genes that are significantly affected downstream of the CRISPR perturbations (https://gwps.wi.mit.edu/)

Group co-regulated genes and co-functional perturbations into groups of genes based on a chosen clustering or dimension reduction (PCA) + clustering algorithm.

Perform enrichment of the PoPS scores foe 120 traits in the genes that are in each program against a background set of random genes selected from the pool of all perturbations.

Perform Stratified LD score regression of the genes in the gene program connected to variants by the cS2G method (https://github.com/bulik/ldsc/wiki)