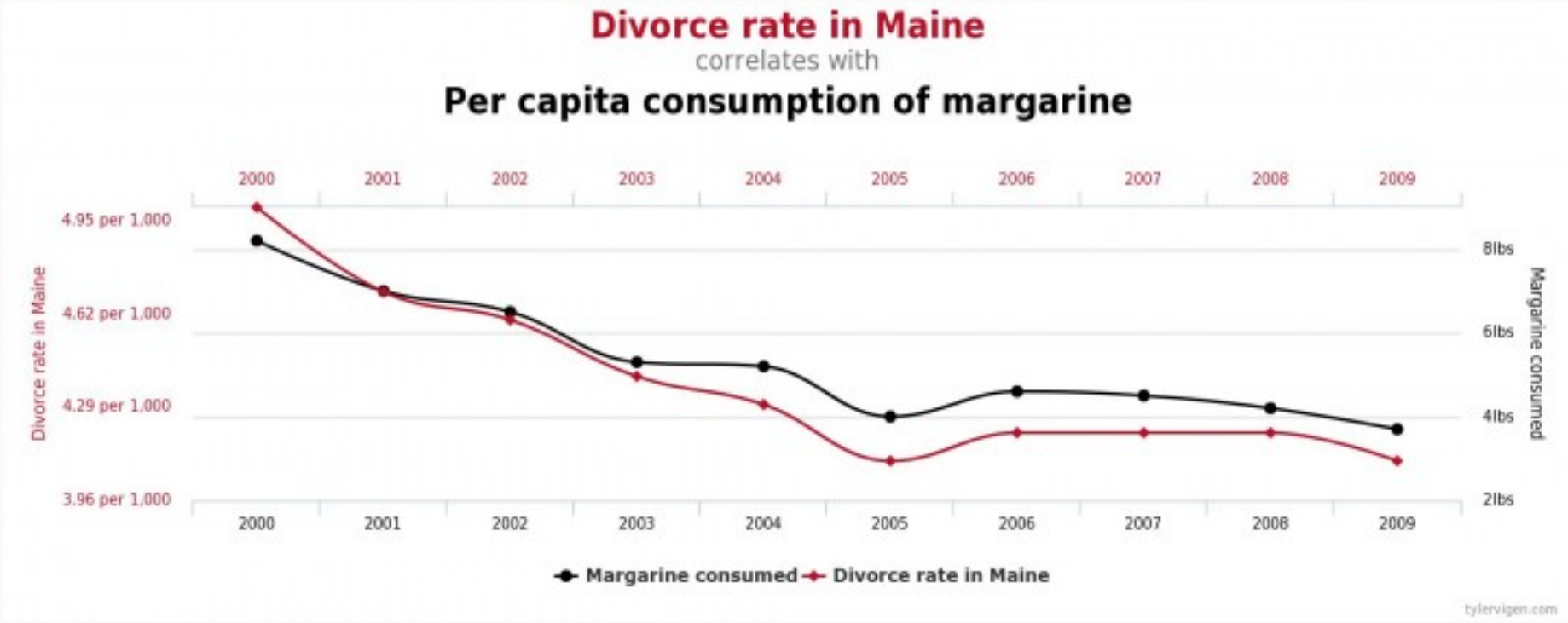
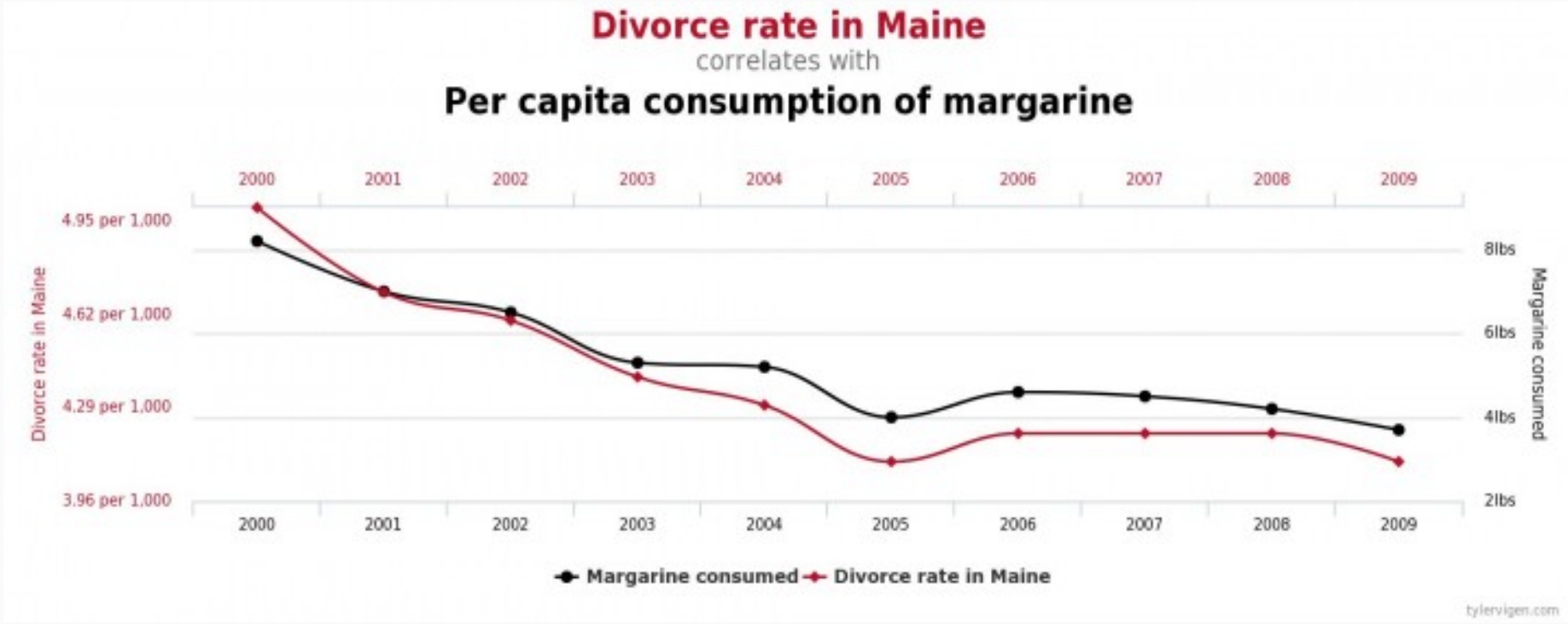
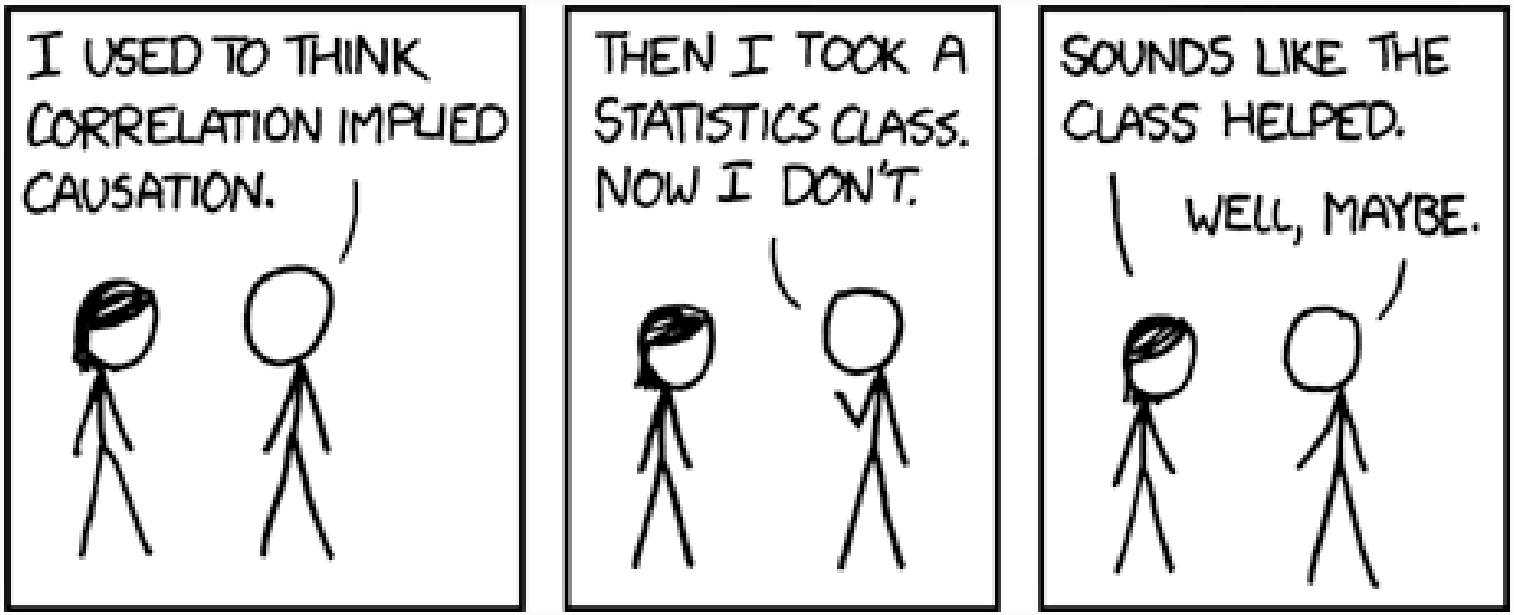
GWAS: from association to causality

Chris Wallace

chr1swallace chr1swallace.github.io cew54@cam.ac.uk


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com

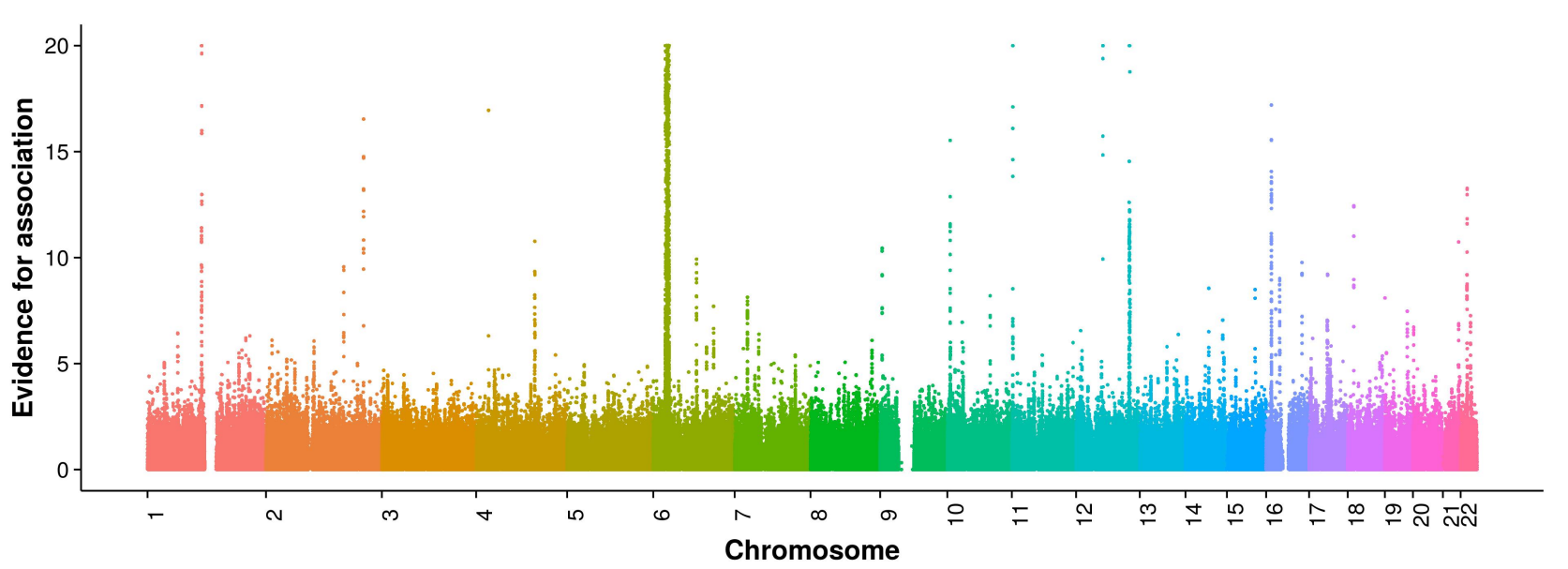

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Outline

- Inferring disease relevant cells
- Fine mapping causal genetic variants
- Causal genes and cells
- 3D genome regulation
- Colocalisation of disease and gene expression variants
- Transcriptome-wide association study

Inferring disease relevant cells

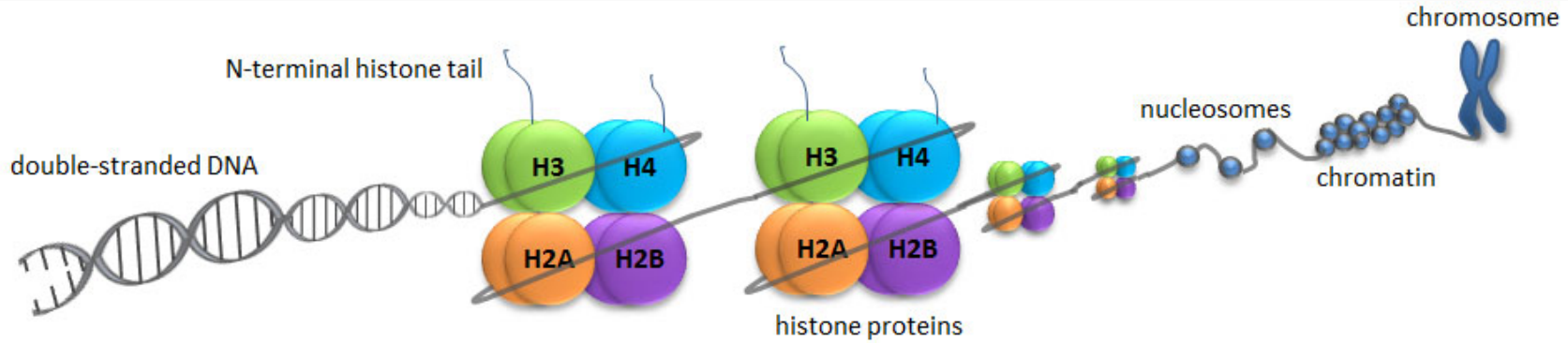
Association does not identify the causal variant



We can learn something from location of *associated* variants

DNA is packed around histone proteins into nucleosomes

Modifications of N-terminal tails of histones vary with accessibility of DNA sequence

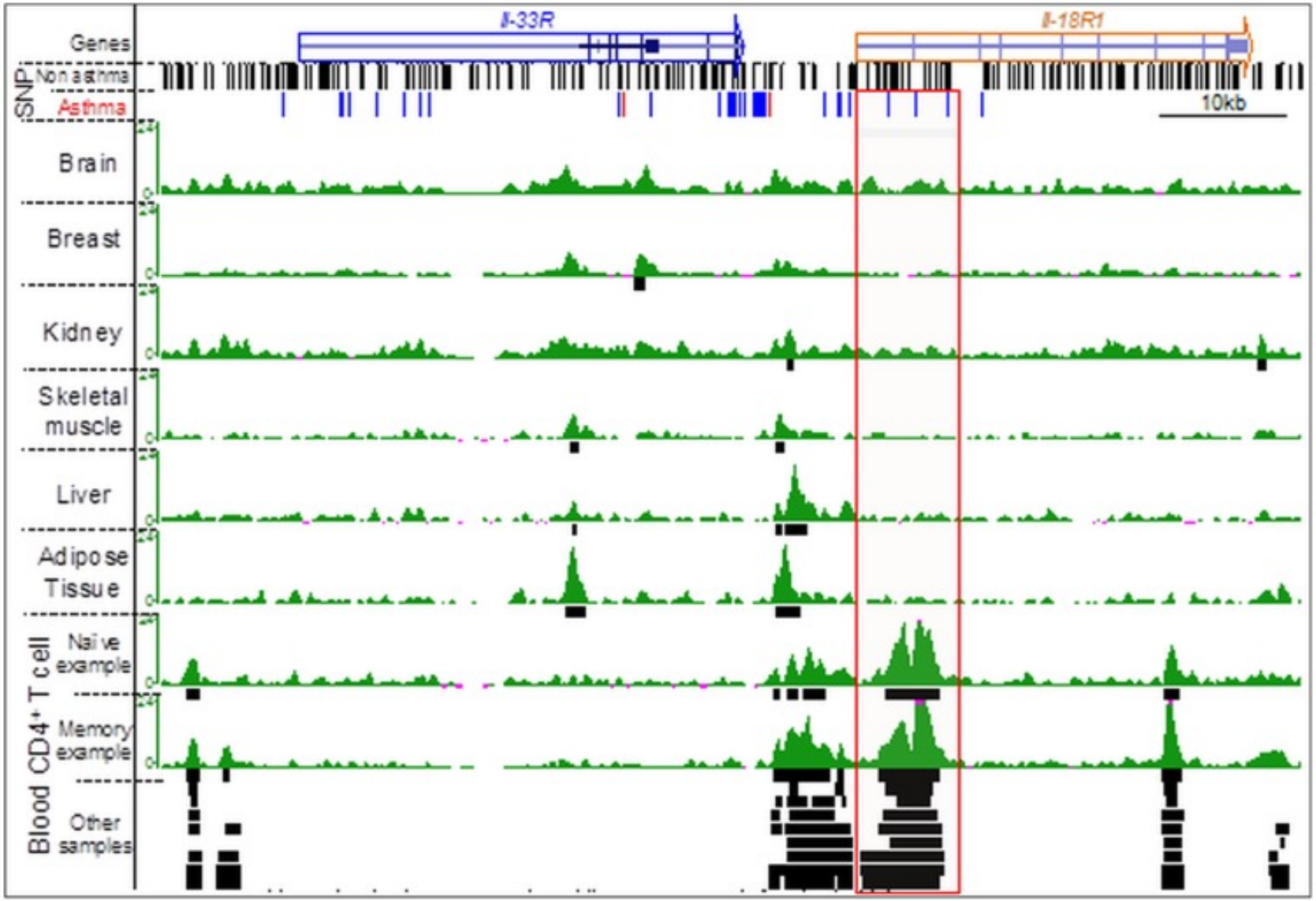


Resources:

ENCODE `https://genome.ucsc.edu/encode`

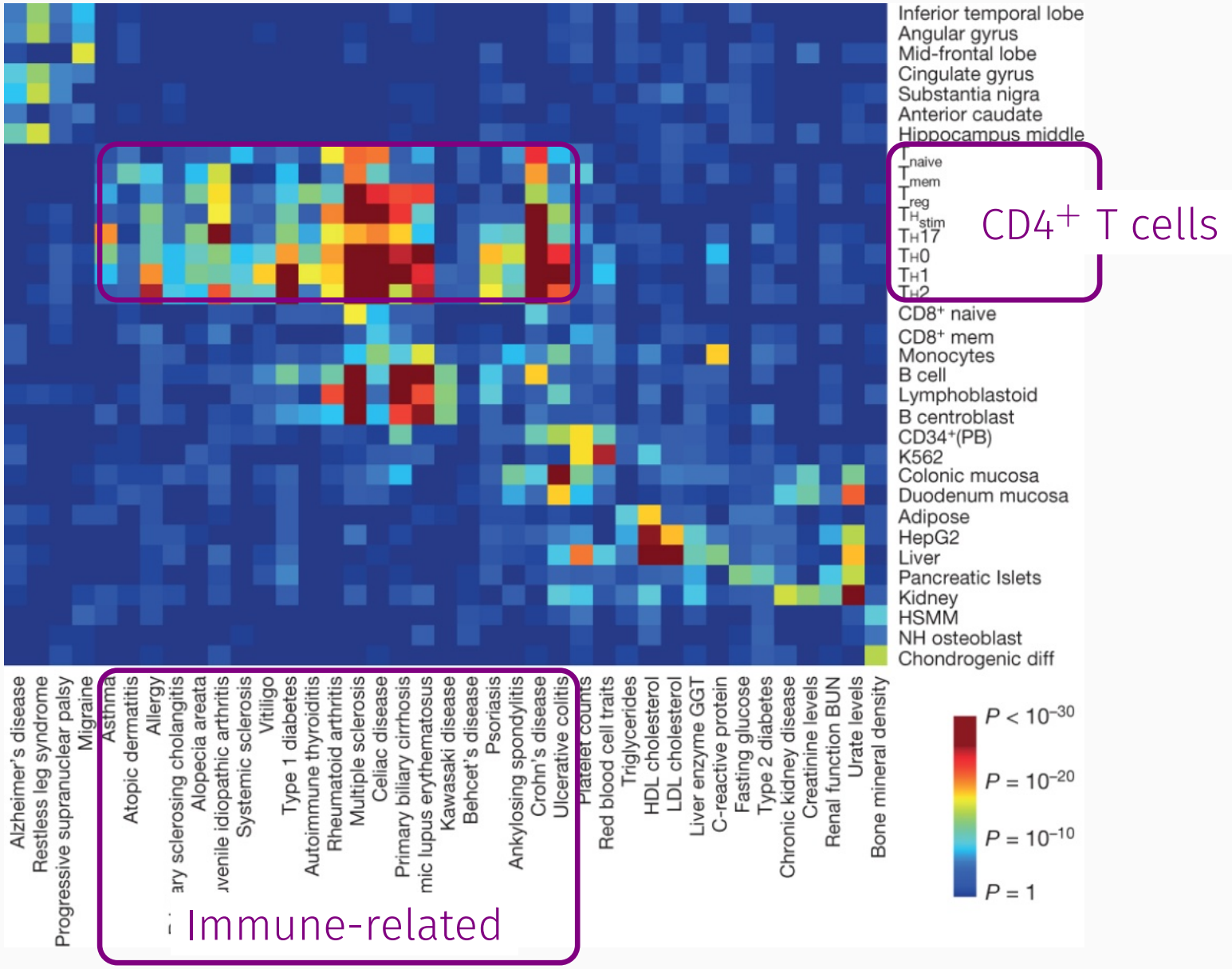
BLUEPRINT `http://www.blueprint-epigenome.eu`

We can learn something from location of *associated* variants



Gerasimova et al, PLOS One 2013

Enrichment of enhancer marks links cells to disease



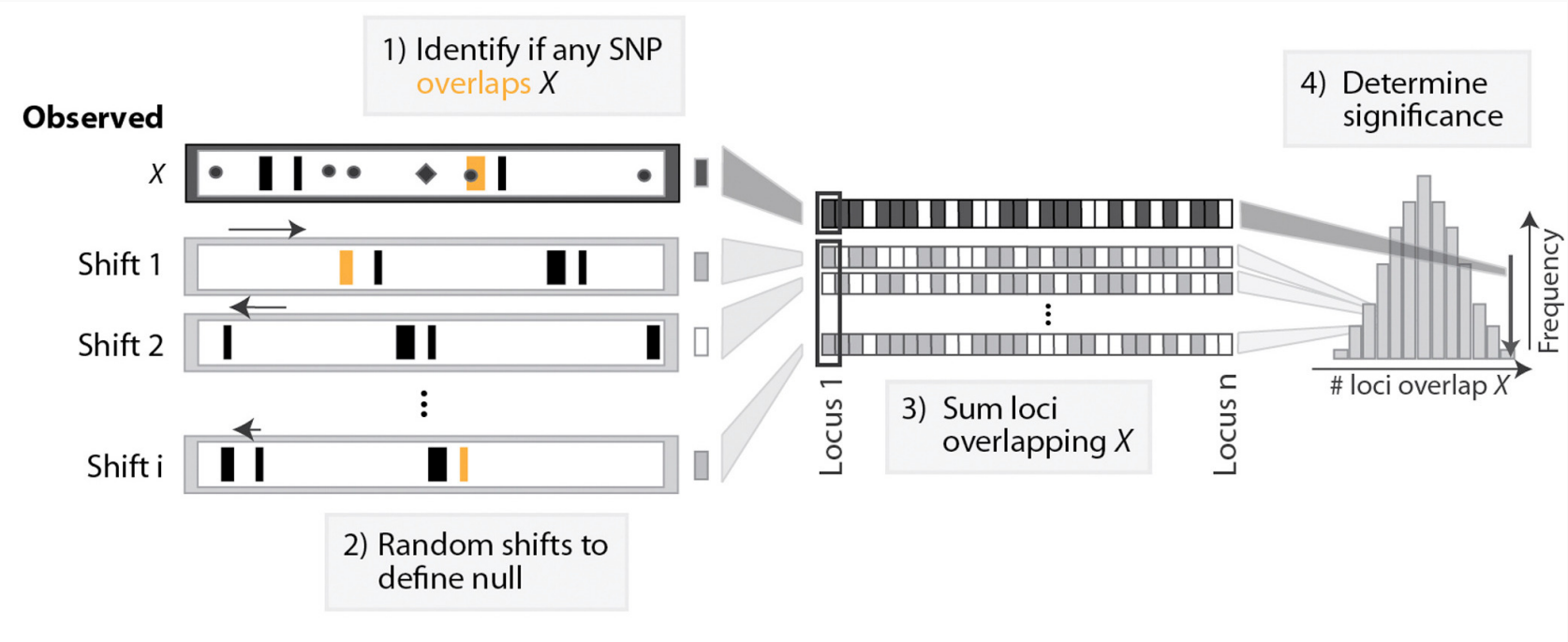
Farh et al., Nature 2014

LD (again)

SNPs are not independent (LD). Chromatin marks spatially cluster.

Ignoring LD makes p values more significant - but is still wrong.

Methods to account for LD maintain the structure of SNP distances when computing null distribution of test statistics.

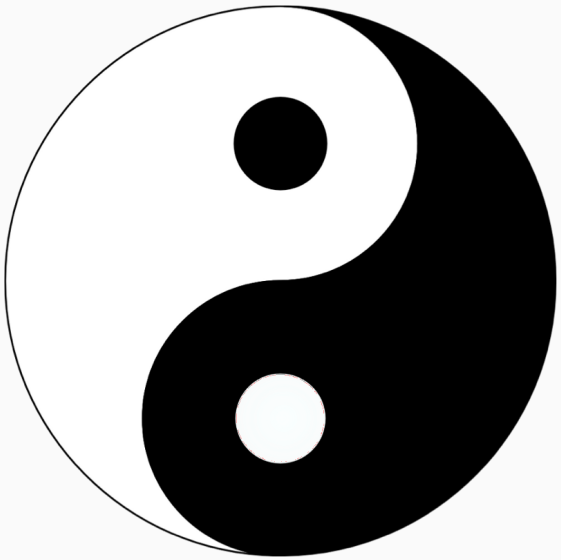


"GoSHIFTER": Trynka et al Am J Hum Genet 2015

Fine mapping causal genetic variants

LD as friend and foe

LD allows us to do GWAS - can use ~500,000 SNPs to scan the whole genome



LD stops us finding causal variants - can't distinguish between variants in strong LD using statistics alone

Fine mapping causal variants

Which variants are causal, which are just “passengers”?

Fine mapping causal variants

Which variants are causal, which are just “passengers”?

A hard statistical problem due to large number of predictors

- ∼1000 correlated predictors
  (genetic variants in linkage disequilibrium)
- Number of models grows exponentially with number of causal variants

Simplifying assumptions:

- Single causal variant: credible set of SNPs
- Uncorrelated causal variants: stepwise regression + credible set for each “hit”

Simplified example of causal variant search

5 SNPs: A, B, C, D, E

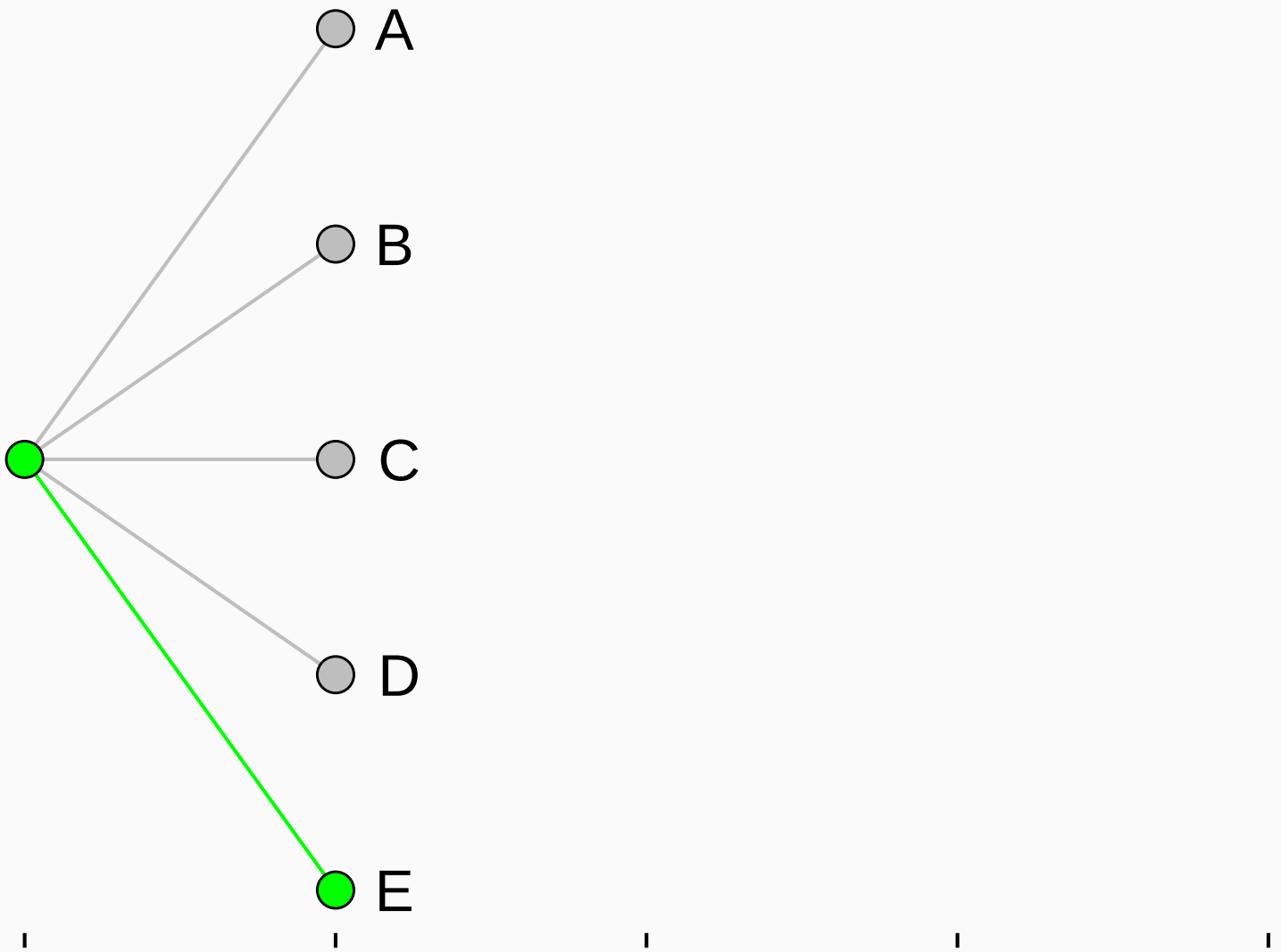
Consider ”models”, combinations of SNPs:

- A
- B + C

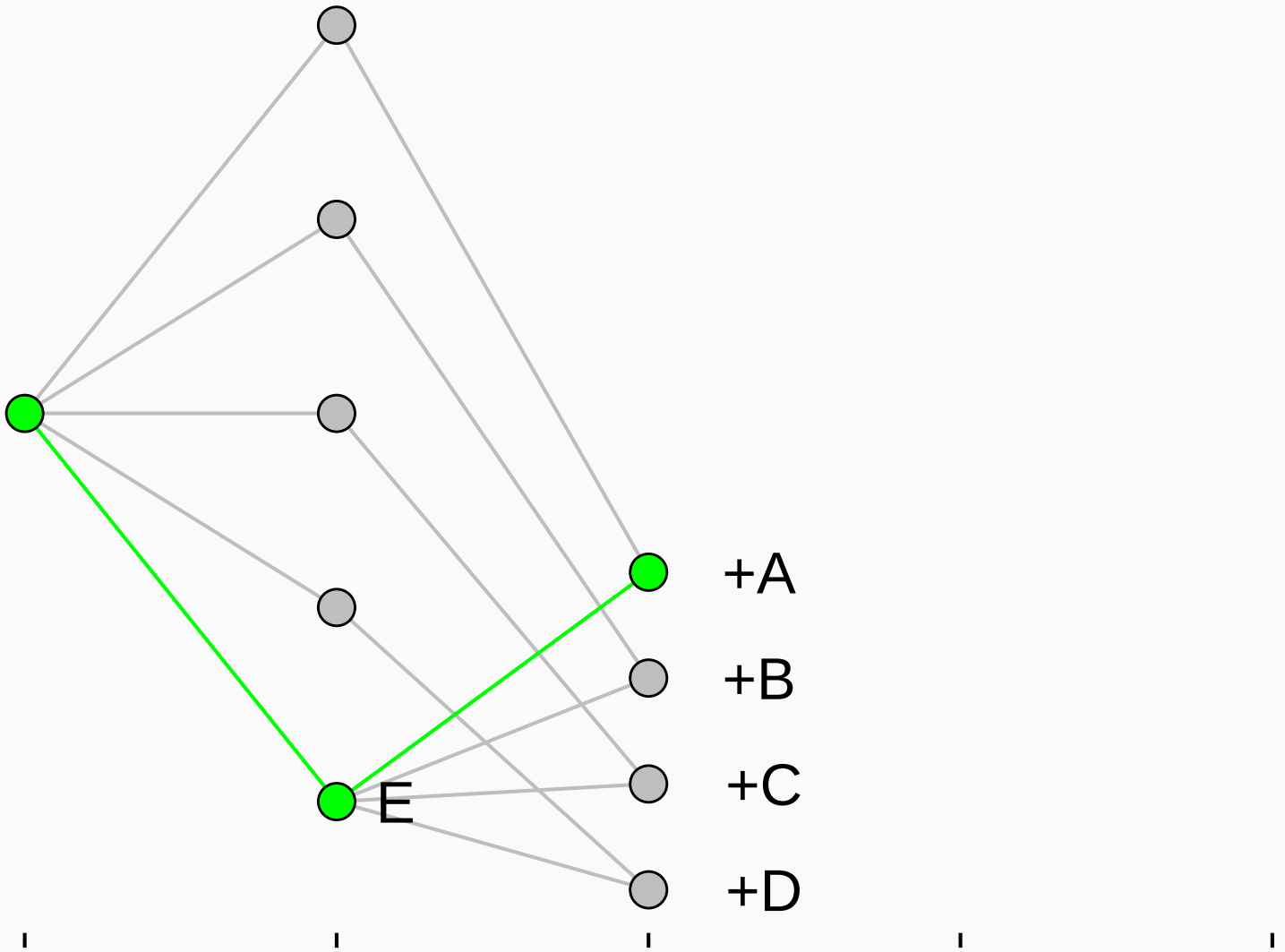
There are 32 possible models. How can we find the ”best” one?

1: start with no SNPs

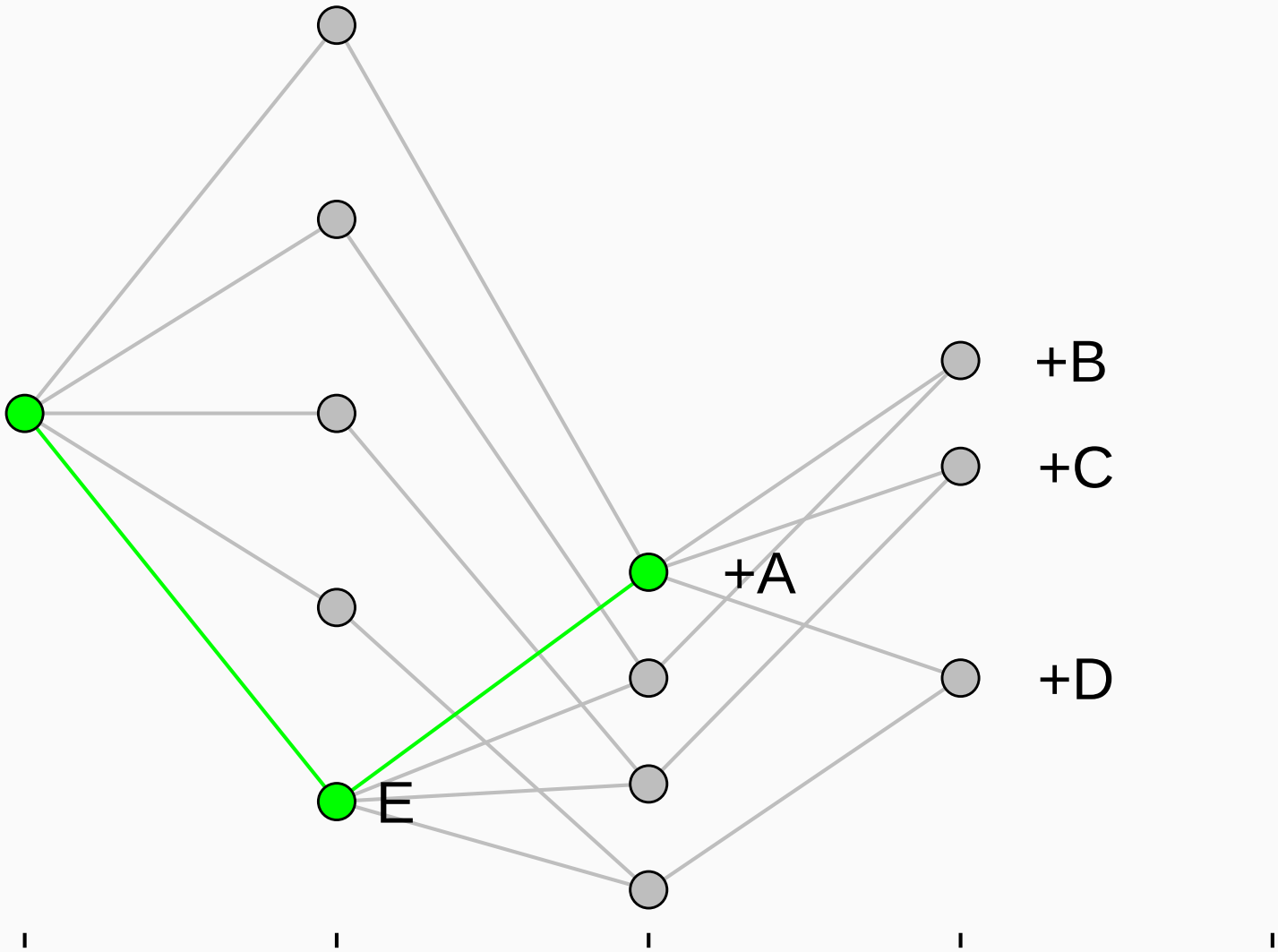
2: best model assuming single causal variant



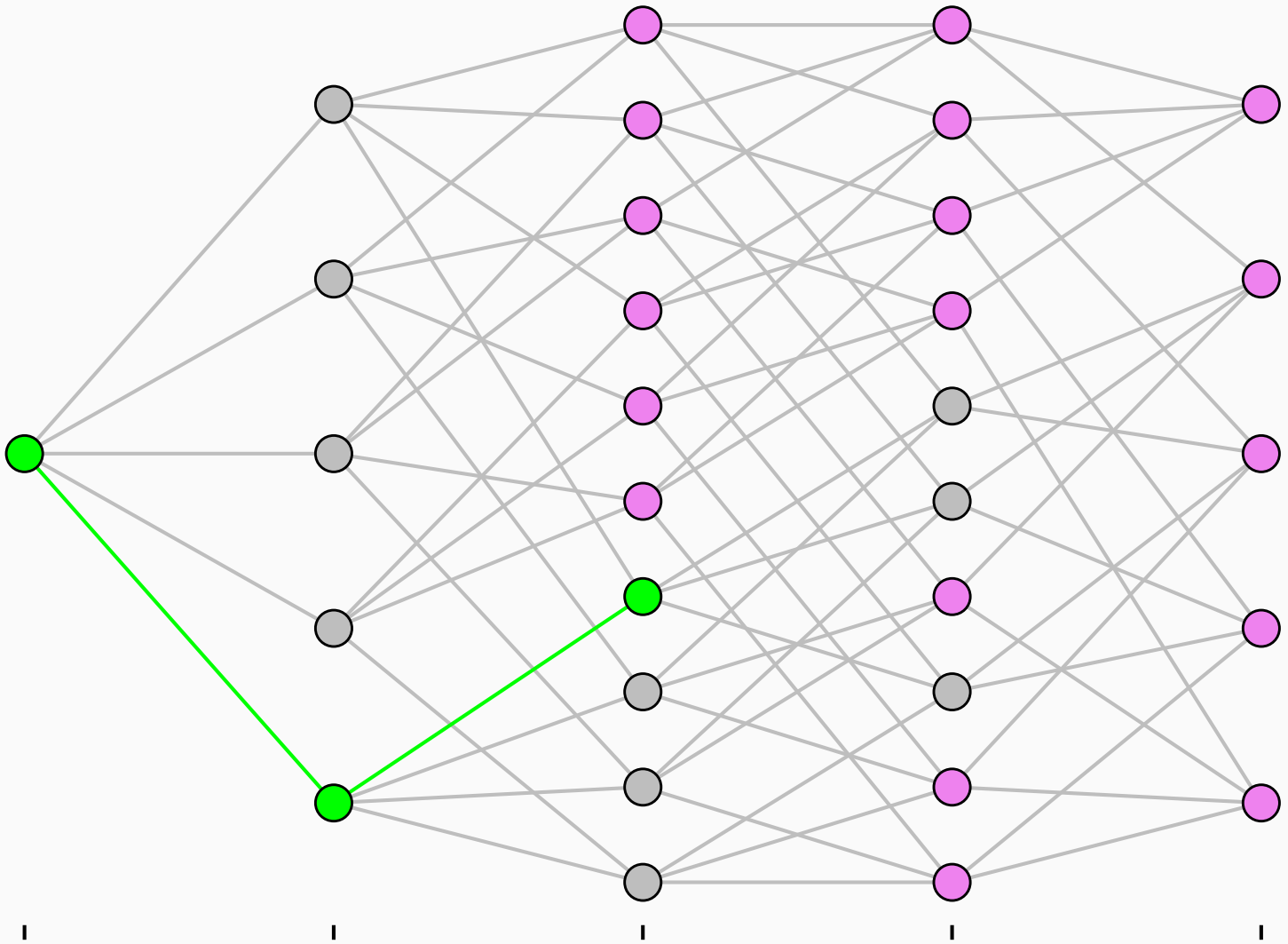
3: step forward into two SNP models



4: consider three SNP models



We only explored a subset of models



Alternative: regularized regression

OLS regression

$$\text{Objective function:} \quad \sum_i (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$$

Stepwise methods find one solution, adding each additional SNP/predictor at each step according to a significance test.

Lasso regression

Adds an L1 penalty to the objective function and finds single solution, with most elements of β "shrunk" to 0.

$$\text{Objective function:} \quad \sum_i (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda \sum_j |\beta_j|$$

λ typically chosen by cross validation

Alternative (2): Joint modelling

Bayes' rule: calculate

$$P(\text{model}|\text{Data}) \propto P(\text{Data}|\text{model})P(\text{model})$$

Advantages: no longer picking one best model
can estimate relative chance of each being true
enforce parsimony - $P(\text{model})$ smaller for larger models

Disadvantages: model space is huge!

Number of SNPs in model	Number of possible models
0	1
1	1,000
2	1000 * 999 / 2 = 499,500
3	1000 * 999 * 998 / (2 * 3) = 166,167,000
⋮	

Joint modelling under single causal variable assumption

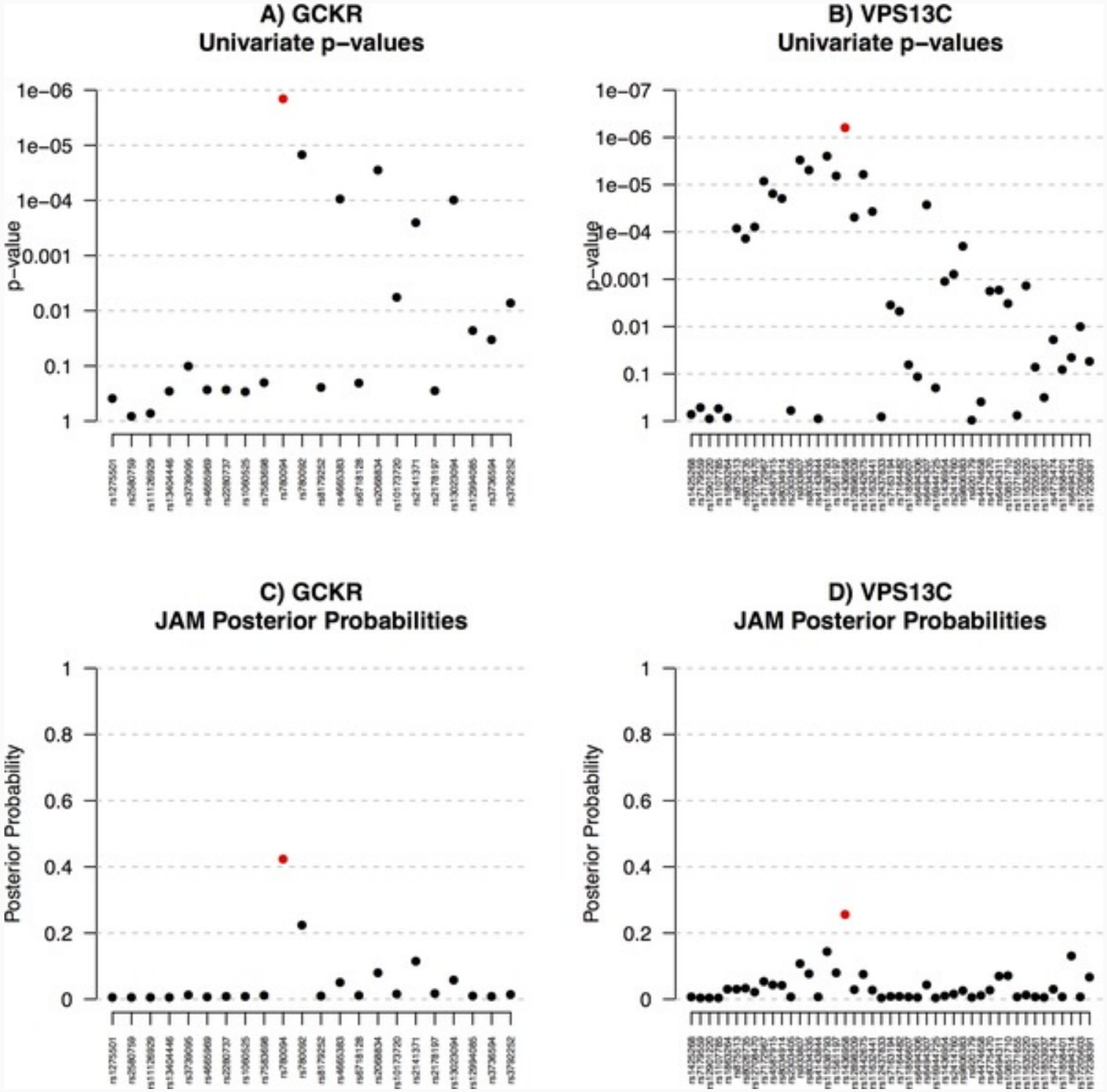
Consider n SNPs.

Assume all 1 SNP models equally likely with Probability π_1 (or not - if you have external information)

Then no SNP model has Probability $\pi_0 = 1 - n\pi_1$

$$P(\text{SNP i causal}) = \frac{P(\text{Data}|\text{SNP i causal})\pi_1}{P(\text{Data}|\text{no SNP causal})\pi_0 + \sum_j P(\text{Data}|\text{SNP } j \text{ causal})\pi_1}$$

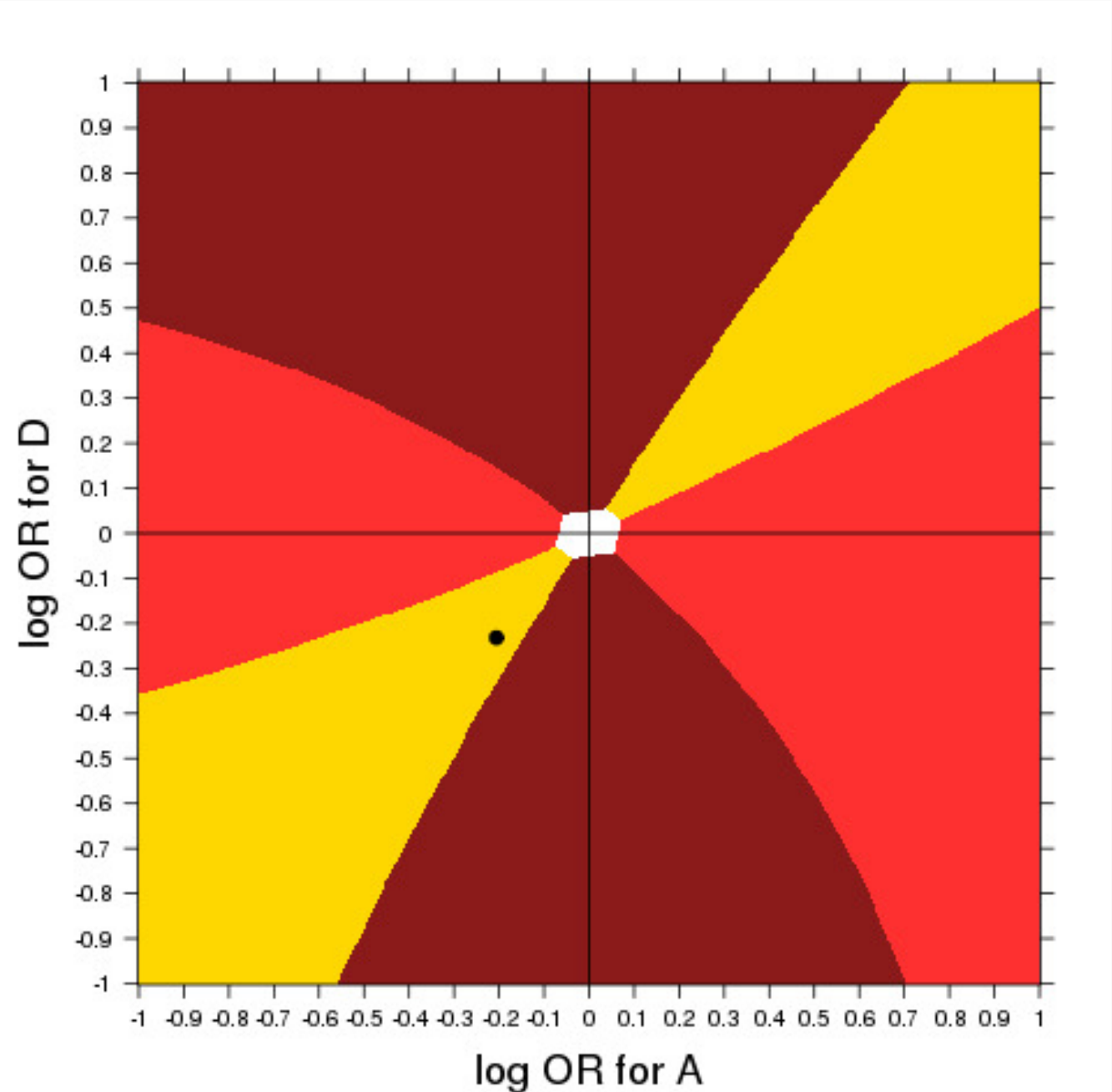
Can calculate these probabilities via Bayes factors. Wakefield (2009)[3] gives methods to calculate approximate *Bayes factors from GWAS p values*


A) GCKR
Univariate p−values
p−value
1e−06
1e−05
1e−04
0.001
0.01
0.1
1
B) VPS13C
Univariate p−values
p−value
1e−07
1e−06
1e−05
1e−04
0.001
0.01
0.1
1
C) GCKR
JAM Posterior Probabilities
Posterior Probability
0
0.2
0.4
0.6
0.8
1
D) VPS13C
JAM Posterior Probabilities
Posterior Probability
0
0.2
0.4
0.6
0.8
1

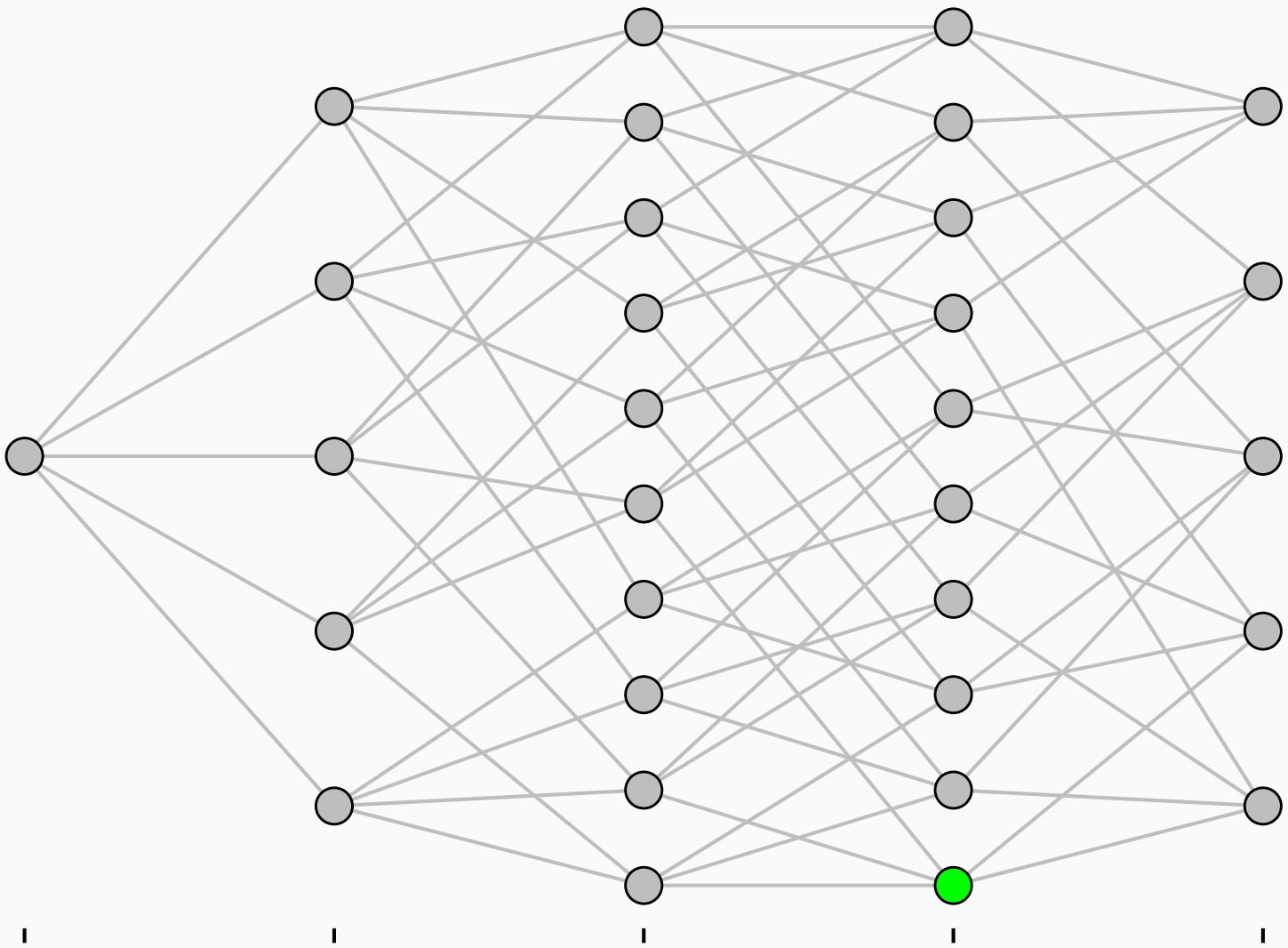
The first SNP might not be (close to) causal

	rs2104286	A: rs12722496	D: rs56382813
p value	7.61E-23	1.16E-05	6.45E-18
r^2			
rs2104286	1.00	**0.33**	**0.31**
A: rs12722496	**0.33**	1.00	0.08
D: rs56382813	**0.31**	0.08	1.00

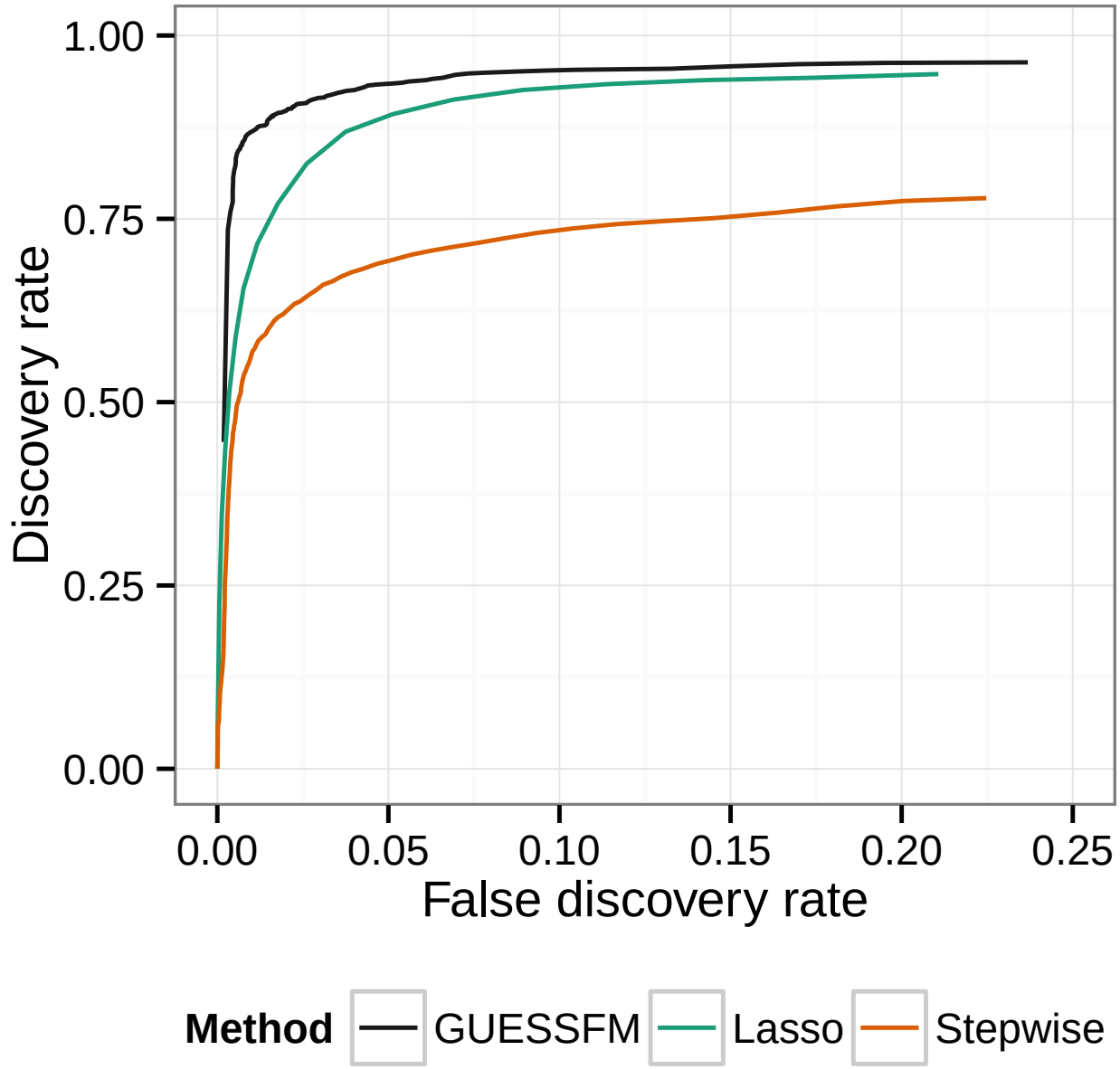
The first SNP might not be (close to) causal

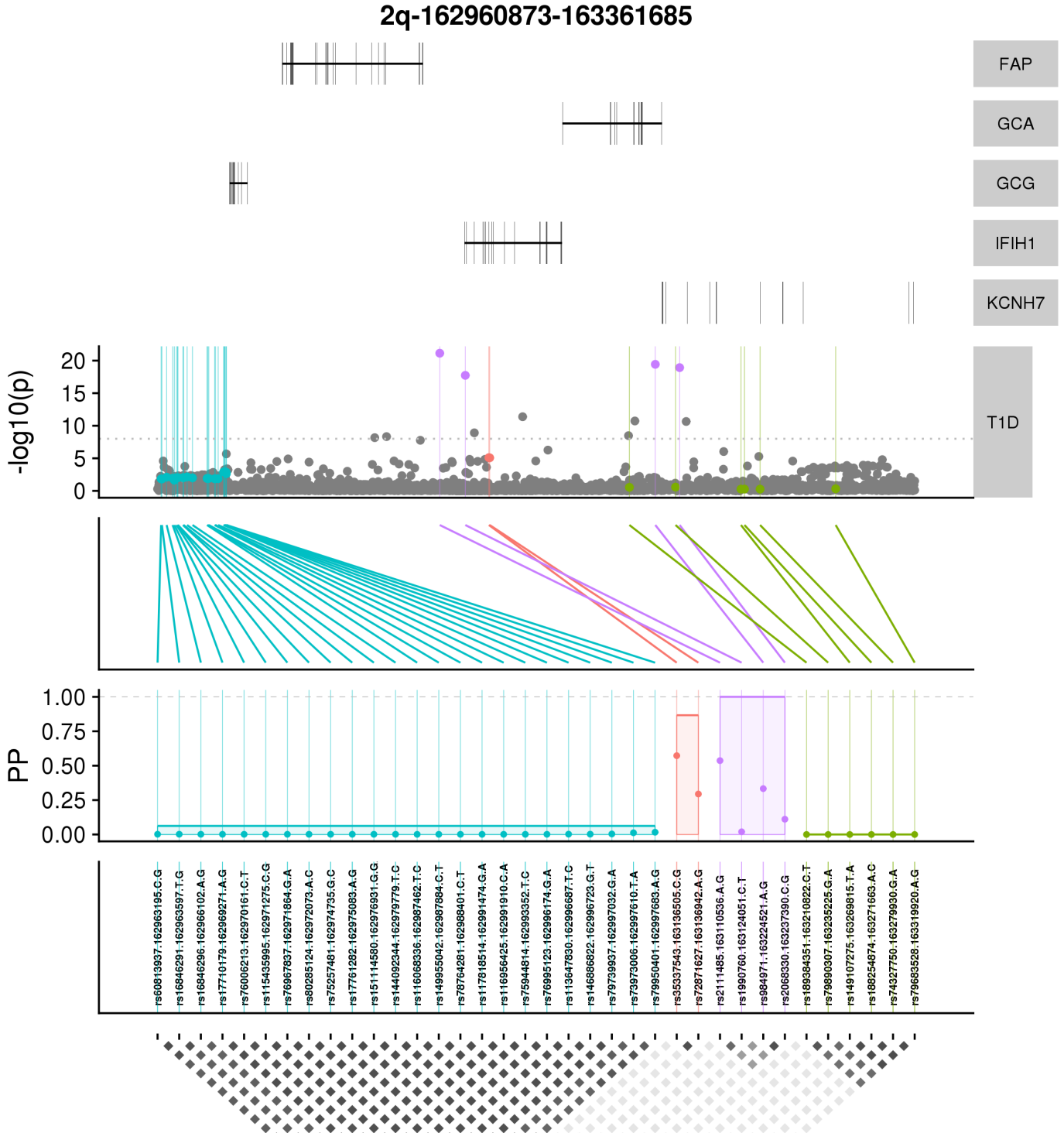


Alternative (3): joint modelling with ‘stochastic search’



Better recovery of “causal” variants in simulated data




2q-162960873-163361685
FAP
GCA
GCG
IFIH1
KCNH7
T1D
-log10(p)
20
15
10
5
0
PP
1.00
0.75
0.50
0.25
0.00
rs60813937.162963195.C.G
rs16846291.162963597.T.G
rs16846296.162966102.A.G
rs177710179.162969271.A.G
rs76006213.162970161.C.T
rs115435995.162971275.C.G
rs76967837.162971864.G.A
rs80285124.162972073.A.C
rs75257481.162974735.G.C
rs17761282.162975083.A.G
rs151114580.162976931.G.G
rs144092344.162979779.T.C
rs116068336.162987462.T.C
rs149955042.162987884.C.T
rs78764281.162988401.C.T
rs117818514.162991474.G.A
rs116956425.162991910.C.A
rs75944814.162993352.T.C
rs76995123.162996174.G.A
rs113647830.162996687.T.C
rs146886822.162996723.G.T
rs79739937.162997032.G.A
rs73973006.162997610.T.A
rs79950401.162997683.A.G
rs35337543.163136505.C.G
rs72871627.163136942.A.G
rs2111485.163110536.A.G
rs1990760.163124051.C.T
rs984971.163224521.A.G
rs2068330.163237390.C.G
rs189384351.163210822.C.T
rs79890307.163235225.G.A
rs149107275.163269815.T.A
rs188254874.163271663.A.C
rs74327750.163279930.G.A
rs79683528.163319920.A.G

Summary of causal variant identification

Stepwise regression is quick, but untrustworthy.
and not just in genetics!

Lasso is better, but focused on prediction, rather than inference

Modelling the joint data is better, but requires simplifying assumptions or specialist software

Stochastic search (via specialist software) is best, but computationally most intensive

Summary of causal variant identification

Stepwise regression is quick, but untrustworthy.
and not just in genetics!

Lasso is better, but focused on prediction, rather than inference

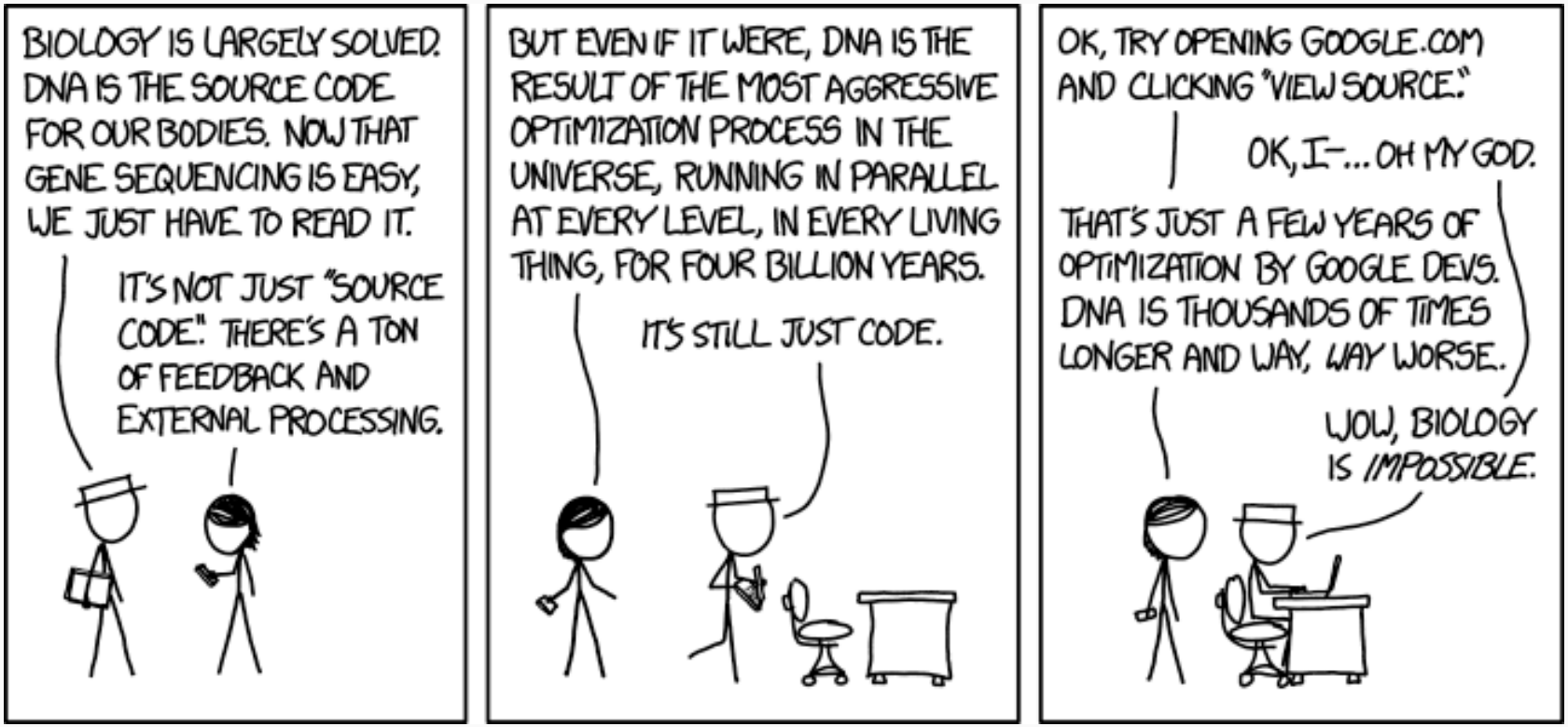
Modelling the joint data is better, but requires simplifying assumptions or specialist software

Stochastic search (via specialist software) is best, but computationally most intensive

---

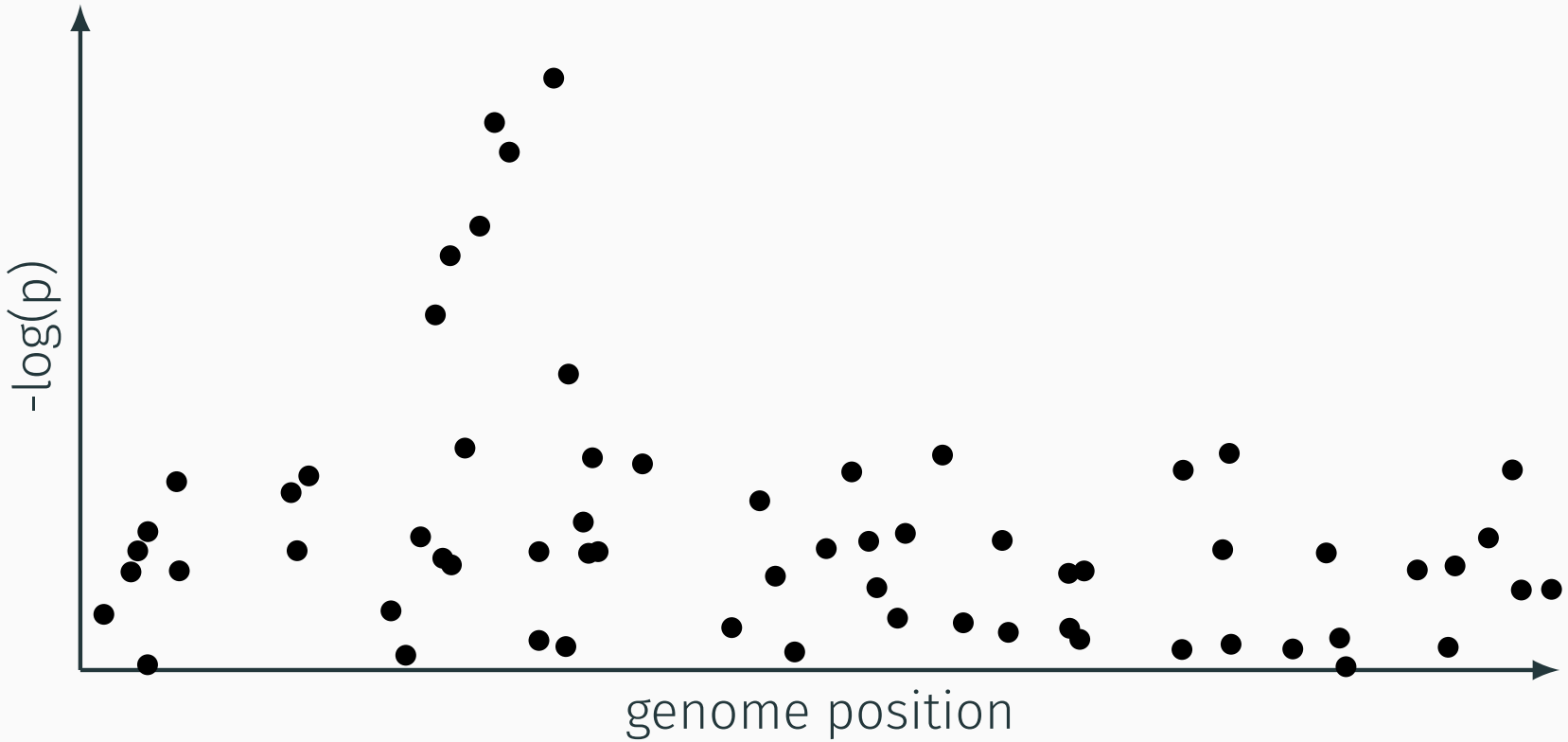
How does knowing causal variants help anyway?

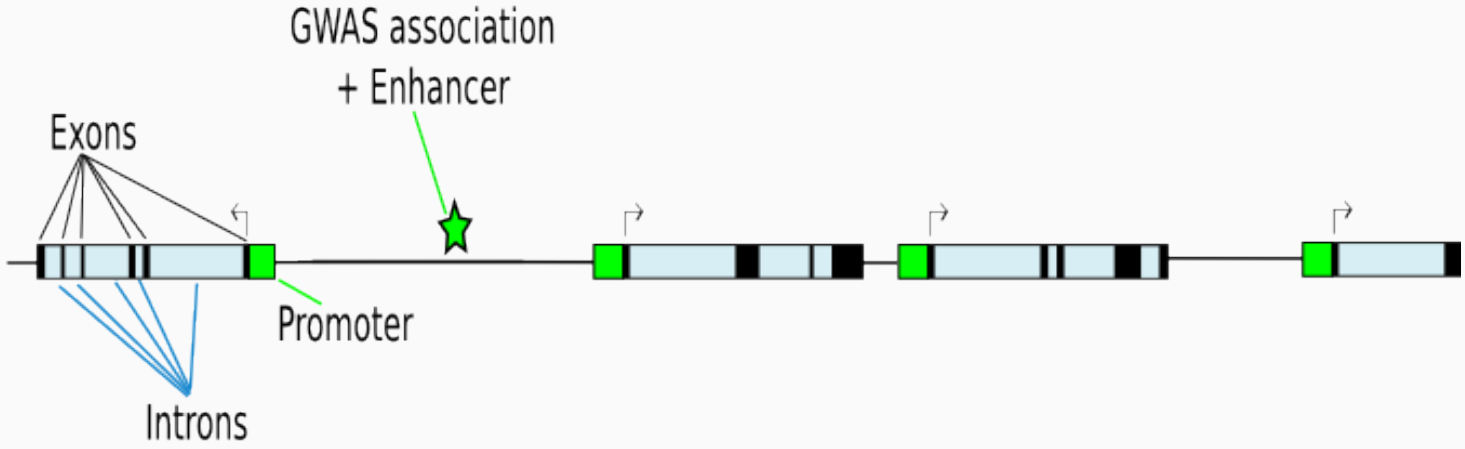
Diversion: biology is complicated



Causal genes and cells

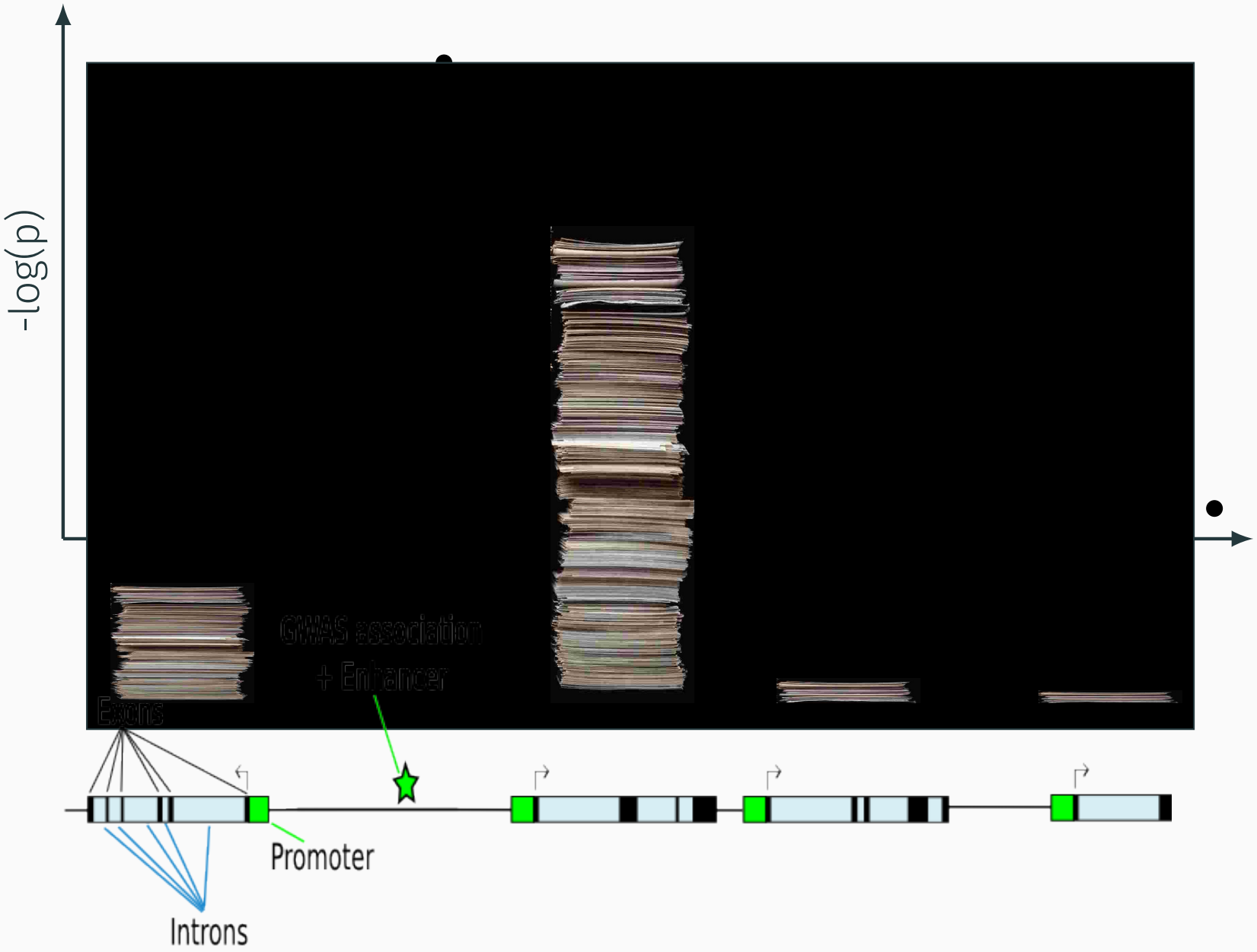
Which *gene* alters disease risk?



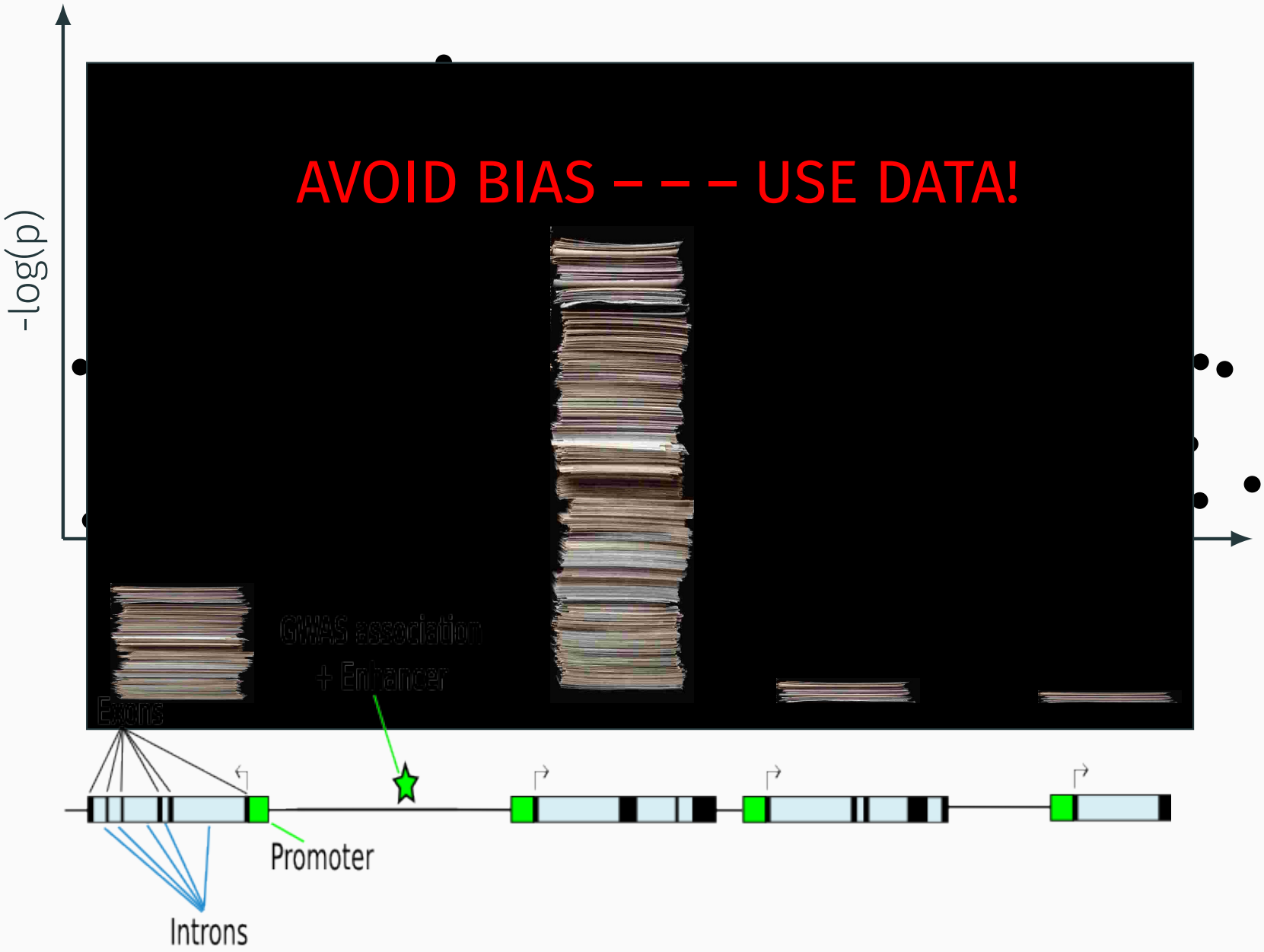


Which *gene* alters disease risk?

Which *gene* alters disease risk?

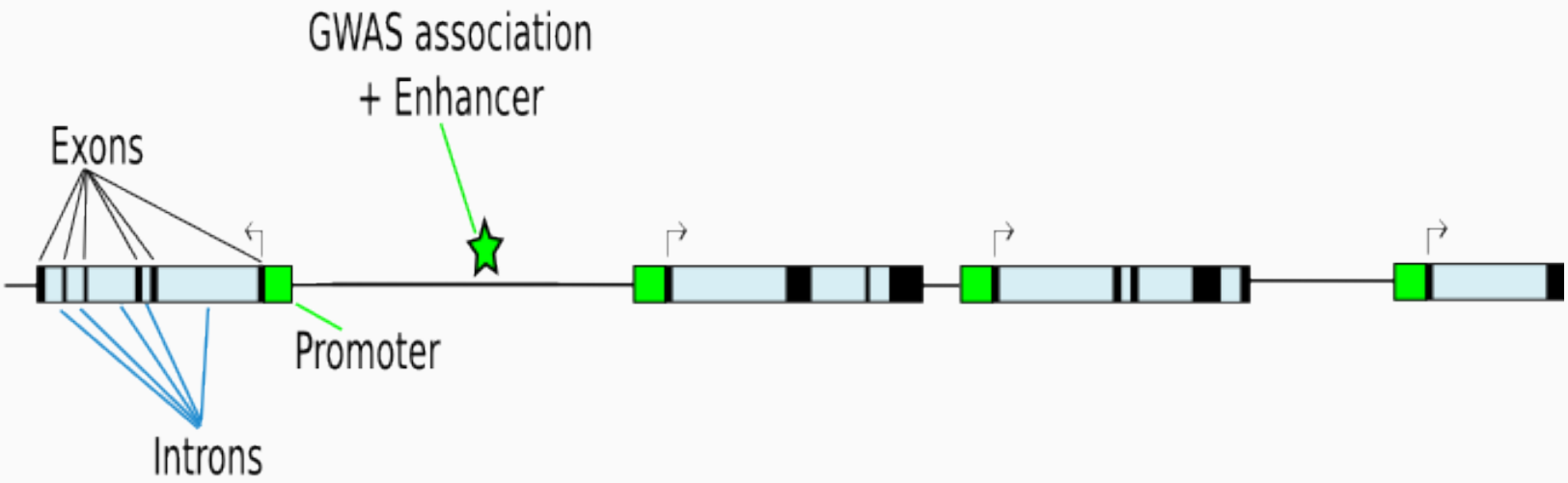


Which *gene* alters disease risk?

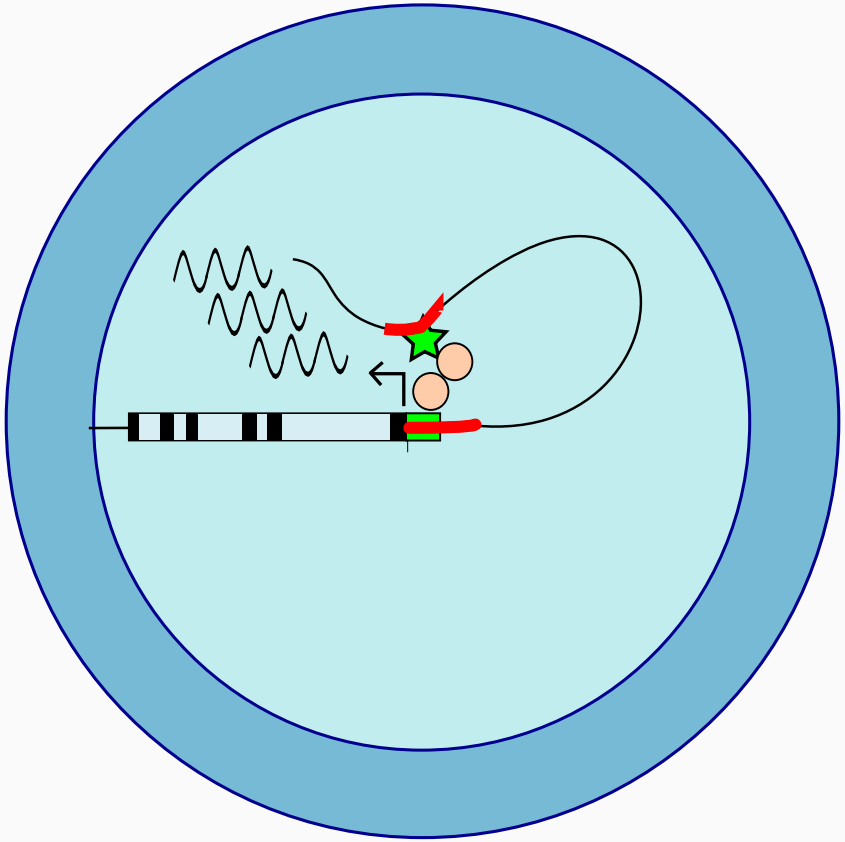


3D genome regulation

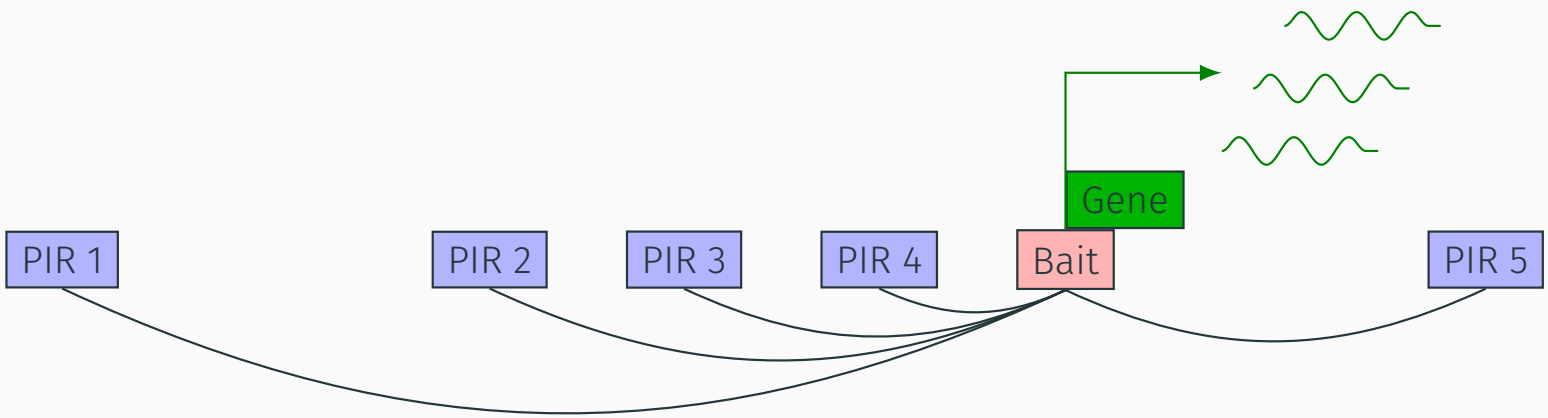
3D folding of DNA in nucleus connects enhancers to promoters



3D folding of DNA in nucleus connects enhancers to promoters

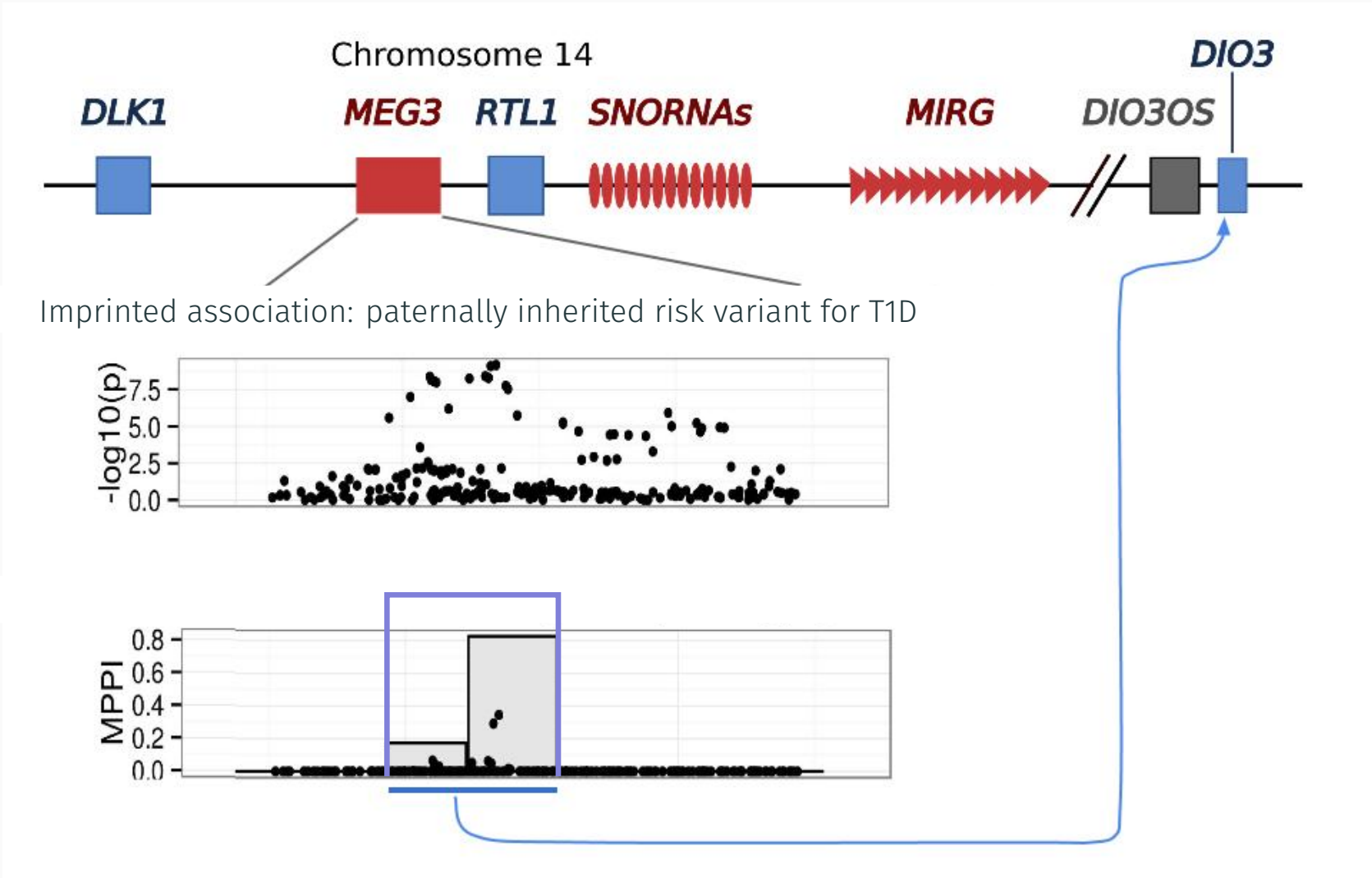


Promoter capture Hi-C



Javierre et al, Cell 2016

Integrate GWAS + PCHi-C to prioritise candidate disease genes



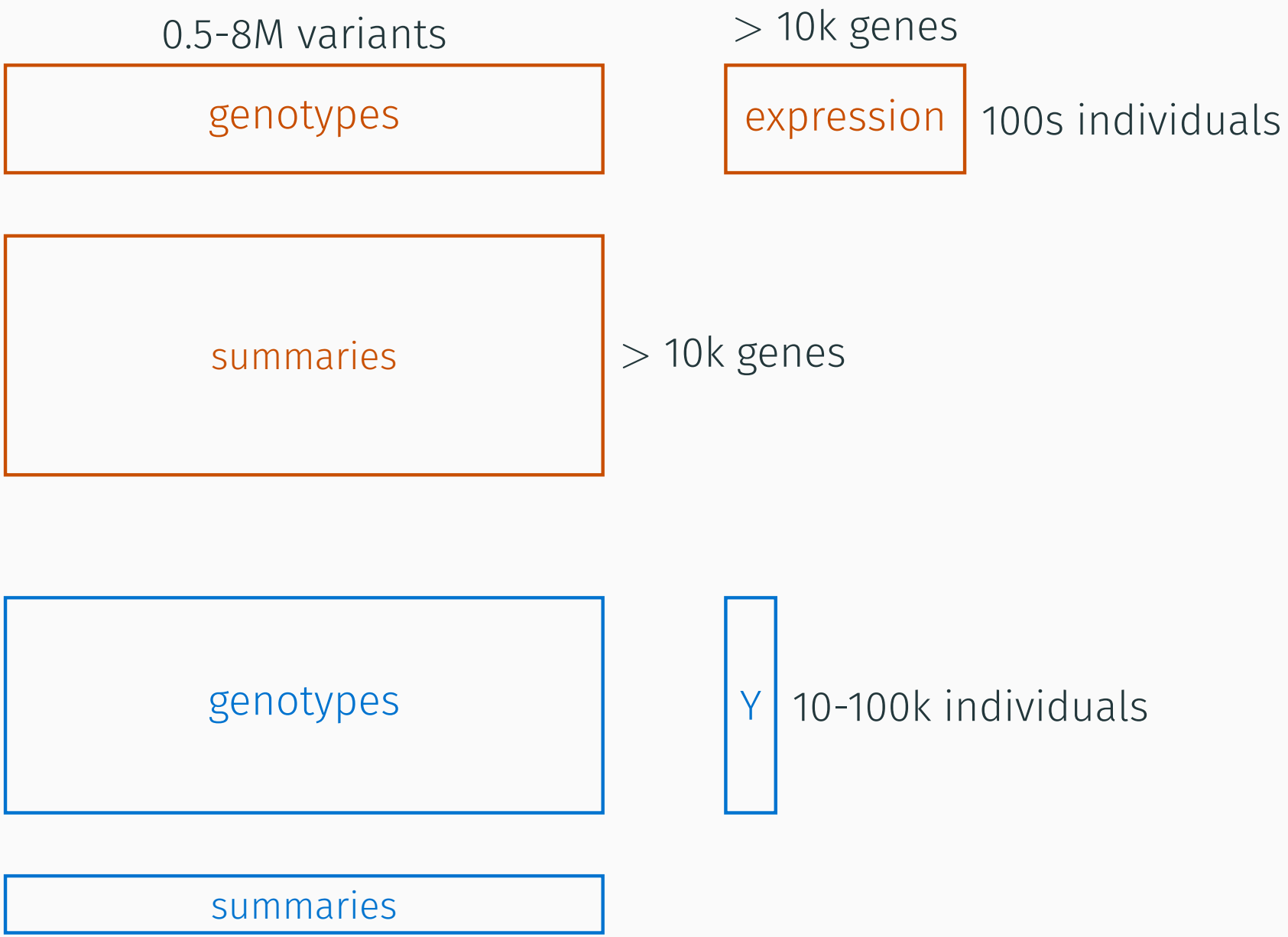
Imprinted association: paternally inherited risk variant for T1D

Aggregate by *Hind*III fragment or set of *Hind*III fragments

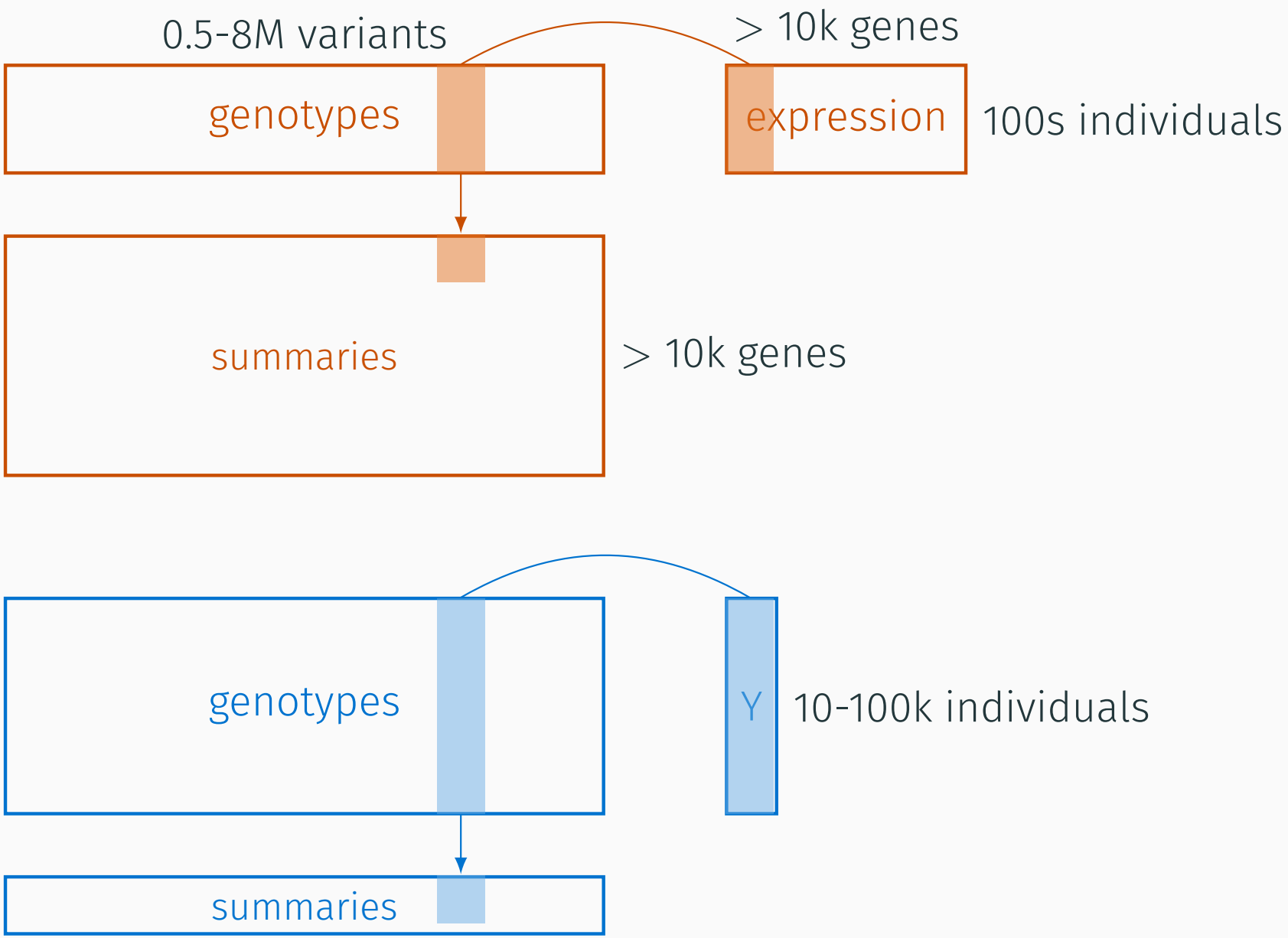
Wallace et al, Nat Genet, 2010

Colocalisation of disease and gene expression variants

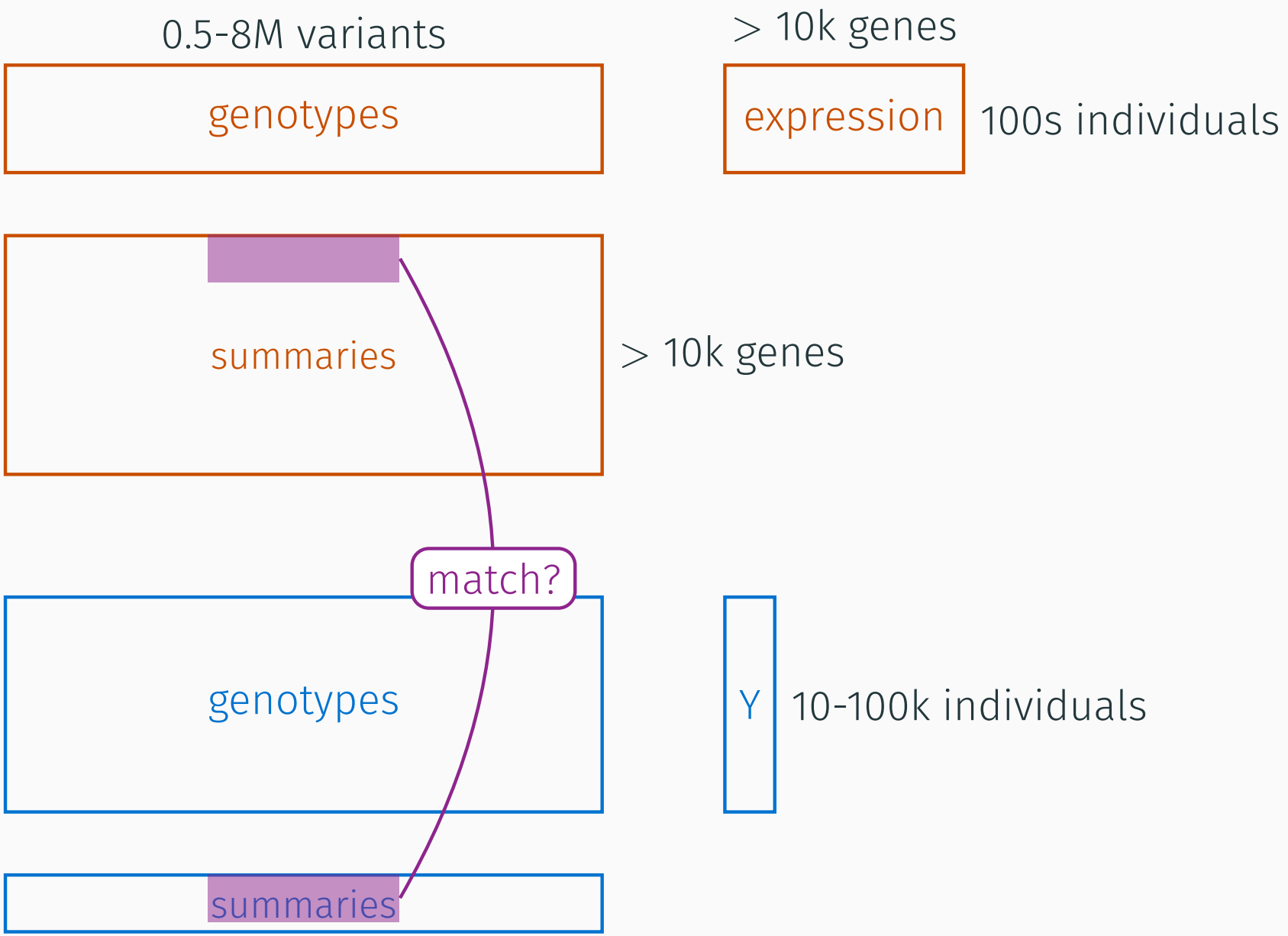
Integrate GWAS and eQTL datasets via colocalisation



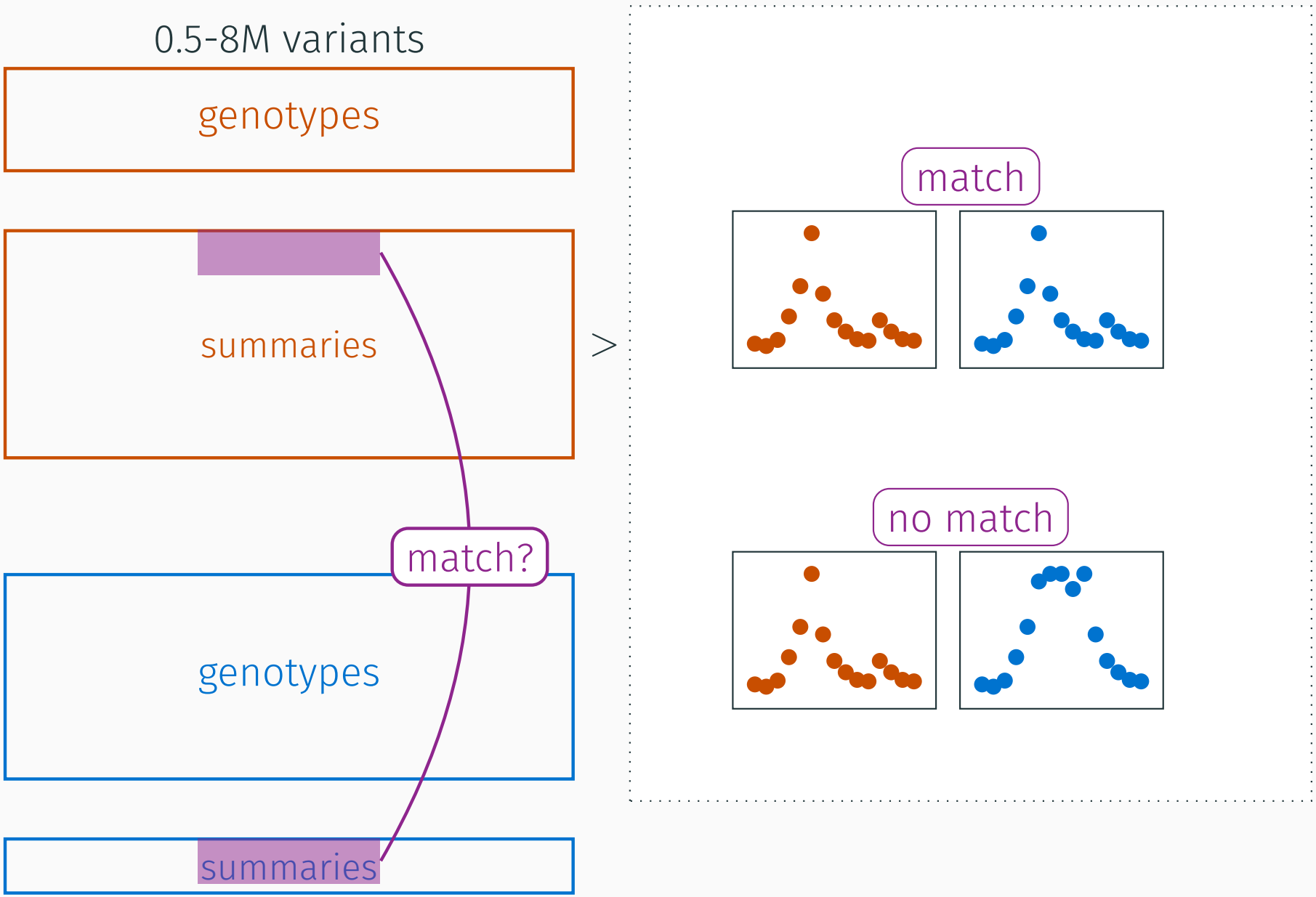
Integrate GWAS and eQTL datasets via colocalisation



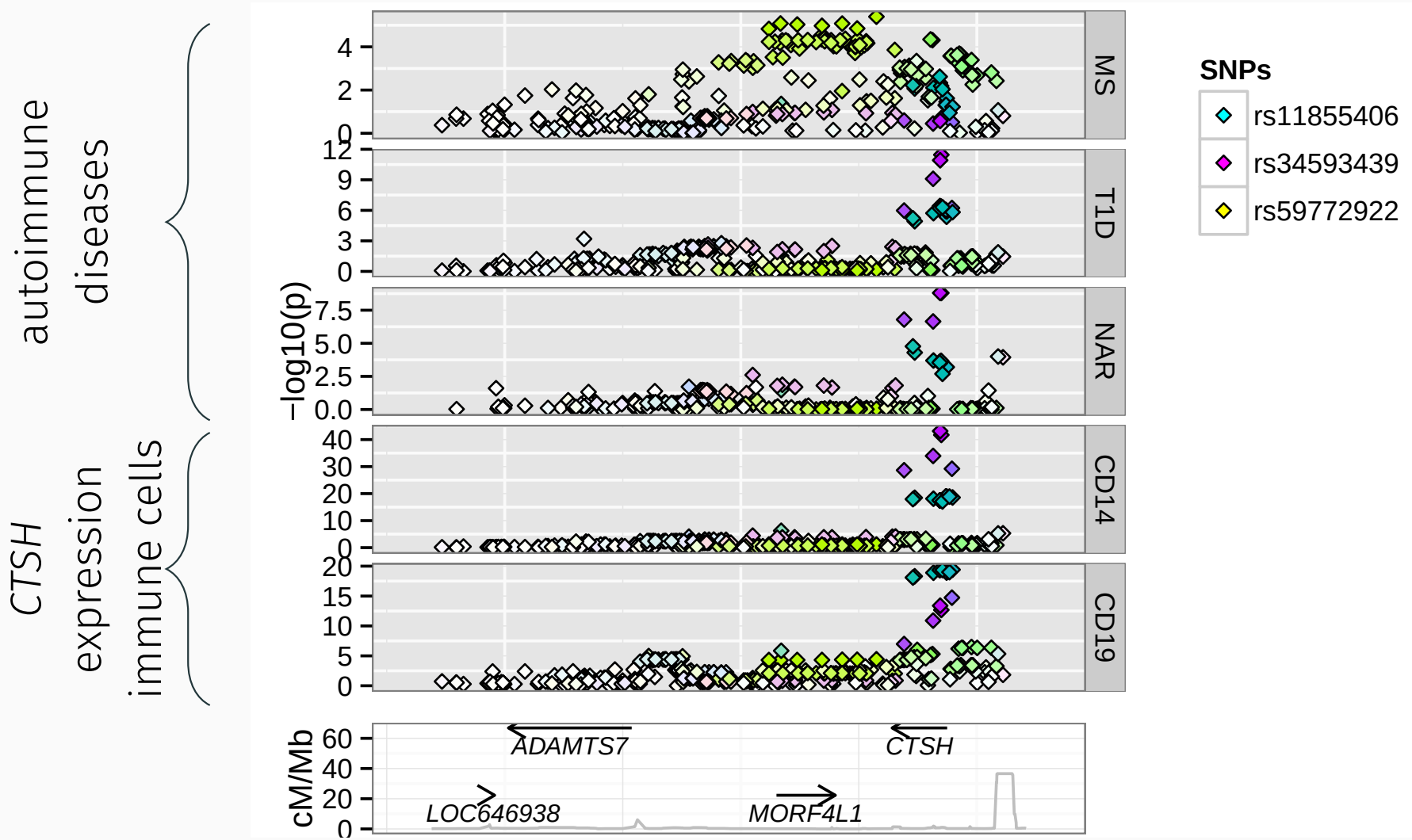
Integrate GWAS and eQTL datasets via colocalisation



Integrate GWAS and eQTL datasets via colocalisation



Colocalisation points to *CTSH* in CD14^{+} monocytes



Guo et al, HMG 2015

Colocalisation anaysis of ten autoimmune diseases with B cell, monocyte, stimulated monocyte eQTLs

+

B cells, monocytes (~ 300 samples)

Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles

Benjamin P Fairfax[1], Seiko Makino[1], Jayachandran Radhakrishnan[1], Katharine Plant[1], Stephen Leslie[2], Alexander Dilthey[3], Peter Ellis[4], Cordelia Langford[4], Fredrik O Vannberg[1,5] & Julian C Knight[1]

Trans-acting genetic vari
expression. Using paired
trans expression quantit
B cells, respectively. Add
disease locus, with *IP6K*
suggesting roles for cell
pathogenesis. We also fi
AOAH and *ARHGAP24*
cell populations identifie
disease susceptibility.

Defining the genetic deter
understanding the biologic
iation. This is particularly
variants underlying obser
wide association studies
functional activity of many
the context of relevant cell
state[2–4]. This context-dep

Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression

Benjamin P. Fairfax,* Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, Julian C. Knight*

READ THE FULL ARTICLE ONLINE
http://dx.doi.org/10.1126/science.1246

Cite this article as B. P. Fairfax *et al.*, *Science* **343**, 1246949 (2014).
DOI: 10.1126/science.1246949

Introduction: Many genetic variants associated with common disease susceptibility occur close to immune-related genes in noncoding DNA, suggestive of a regulatory function. The definition of functional variants and the specific genes that they regulate remains challenging and in many cases is unresolved. We hypothesized that a significant proportion of variants, including those implicated in disease, may show activity in a context-specific manner and therefore only be identifiable upon triggering of immune responses.

Methods: We mapped interindividual variation in gene expression as a quantitative trait, defining expression quantitative trait loci (eQTLs). To investigate the effect of innate immune stimuli on eQTLs, we exposed primary CD14$^+$ human monocytes from 432 European volunteers to the inflammatory proxies interferon-γ (IFN-γ) or differing durations (2 or 24 hours) of lipopolysaccharide (LPS). eQTL mapping was performed on a genome-wide basis with an additive linear model. A subset of 228 individuals with expression data available for all experimental conditions enabled cross-treatment comparisons.

Results: Stimulation with LPS or IFN-γ resulted in profound effects across monocyte eQTLs, with hundreds of genes and associated pathways demonstrating context-specific eQTLs dependent on the type and duration of stimulus. Context-specific eQTLs frequently intersected established canonical pathways of monocyte signaling and included key nodal genes and effector molecules. These eQTLs are typically more distal to the transcriptional start site and, in some cases, showed reversal of

on context, paralleling the expression of class II genes. Induced eQTLs were enriched for disease-risk loci with context-specific associations to many putative causal genes including at *ATM*, *IRF8*,

FIGURES IN THE FULL ARTICLE

Fig. 1. Genotype modulates the gene expression response to innate immune stimuli in monocytes.

Fig. 2. Trans-eQTL demonstrate context specificity and identify master regulatory loci after treatment.

Fig. 3. Temporal effects for a stimulus-specific trans-eQTL.

Fig. 4. Cis regulation of IRF2 at rs13149
has profound transcriptional consequenc
in trans.

Fig. 5. Stimulus-specific eQTL and GWAS.

Fig. 6. Examples of context-specific eQTL
informative for disease risk.

SUPPLEMENTARY MATERIALS

Materials and Methods
Figs. S1 to S12

ation with di
signaling.
VB1 expressio

a single–nucleotide polymorphism (rs2275
revealed after 2 hours of LPS stimulation of

Activated monocytes (~ 300 samples)

=

?

candidate causal autoimmune genes

Guo et al., Hum Mol Genet, *2015*

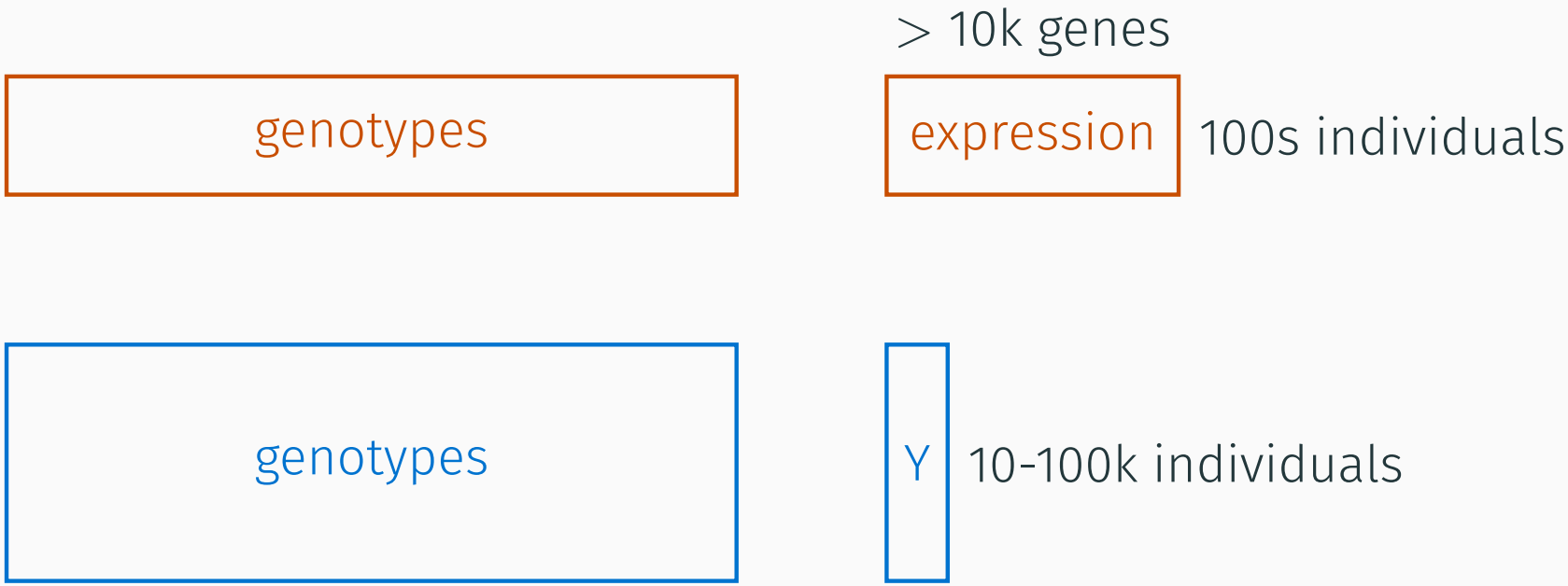
Colocalisation identified six candidate causal autoimmune disease genes

Gene	Disease(s)	Direction
Resting B cells + monocytes		
RGS1	Celiac, MS	-
SYNGR1	Primary Biliary cirrhosis	+
Resting + activated monocytes		
ADAM15	Crohn's	?
CARD19	Crohn's, ulcerative colitis	+
LTBR	Primary Biliary cirrhosis	+
CTSH	T1D, narcolepsy	-

Guo et al., Hum Mol Genet, *2015*

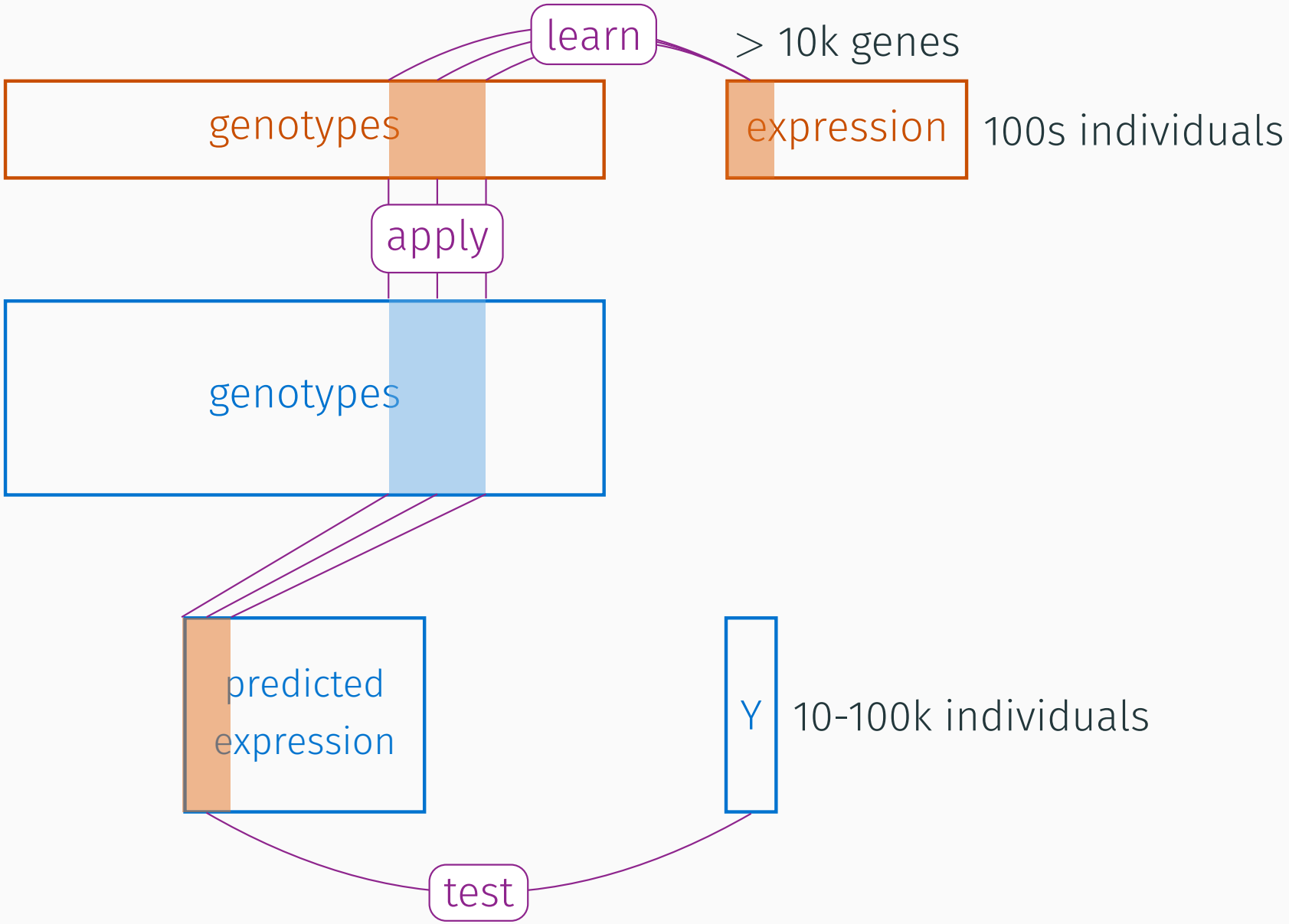
Transcriptome-wide association study

Integrate GWAS and eQTL datasets via TWAS



Gamazon et al, Nat Genet, 2015; Gusev et al, Nat Genetics, 2016

Integrate GWAS and eQTL datasets via TWAS



Gamazon et al, Nat Genet, 2015; Gusev et al, Nat Genetics, 2016

TWAS of T1D with B cell, monocyte, stimulated monocyte eQTLs

+

B cells, monocytes (~ 300 samples)

Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles

Benjamin P Fairfax[1], Seiko Makino[1], Jayachandran Radhakrishnan[1], Katharine Plant[1], Stephen Leslie[2], Alexander Dilthey[3], Peter Ellis[4], Cordelia Langford[4], Fredrik O Vannberg[1,5] & Julian C Knight[1]

Trans-acting genetic vari
expression. Using paired
trans expression quantit
B cells, respectively. Add
disease locus, with *IP6K*
suggesting roles for cell
pathogenesis. We also fi
AOAH and *ARHGAP24* i
cell populations identifie
disease susceptibility.

Defining the genetic deter
understanding the biologic
iation. This is particularly
variants underlying obser
wide association studies
functional activity of many
the context of relevant cell
state[2–4]. This context-dep

Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression

Benjamin P. Fairfax,* Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, Julian C. Knight*

Introduction: Many genetic variants associated with common disease susceptibility occur close to immune-related genes in noncoding DNA, suggestive of a regulatory function. The definition of functional variants and the specific genes that they regulate remains challenging and in many cases is unresolved. We hypothesized that a significant proportion of variants, including those implicated in disease, may show activity in a context-specific manner and therefore only be identifiable upon triggering of immune responses.

Methods: We mapped interindividual variation in gene expression as a quantitative trait, defining expression quantitative trait loci (eQTLs). To investigate the effect of innate immune stimuli on eQTLs, we exposed primary CD14+ human monocytes from 432 European volunteers to the inflammatory proxies interferon-γ (IFN-γ) or differing durations (2 or 24 hours) of lipopolysaccharide (LPS). eQTL mapping was performed on a genome-wide basis with an additive linear model. A subset of 228 individuals with expression data available for all experimental conditions enabled cross-treatment comparisons.

Results: Stimulation with LPS or IFN-γ resulted in profound effects across monocyte eQTLs, with hundreds of genes and associated pathways demonstrating context-specific eQTLs dependent on the type and duration of stimulus. Context-specific eQTLs frequently intersected established canonical pathways of monocyte signaling and included key nodal genes and effector molecules. These eQTLs are typically more distal to the transcriptional start site and, in some cases, showed reversal of

on context, paralleling the expression of class II genes. Induced eQTLs were enriched for disease-risk loci with context-specific associations to many putative causal genes including at *ATM*, *IRF8*,

READ THE FULL ARTICLE ONLINE
http://dx.doi.org/10.1126/science.1246

Cite this article as B. P. Fairfax *et al.*, *Science* **343**, 1246949 (2014).
DOI: 10.1126/science.1246949

FIGURES IN THE FULL ARTICLE

Fig. 1. Genotype modulates the gene expression response to innate immune stimuli in monocytes.

Fig. 2. Trans-eQTL demonstrate context specificity and identify master regulatory loci after treatment.

Fig. 3. Temporal effects for a stimulus-specific trans-eQTL.

Fig. 4. Cis regulation of IRF2 at rs13149
has profound transcriptional consequenc
in trans.

Fig. 5. Stimulus-specific eQTL and GWAS.

Fig. 6. Examples of context-specific eQTL
informative for disease risk.

SUPPLEMENTARY MATERIALS

Materials and Methods
Figs. S1 to S12

signaling.
VB1 expressio
a single–nucleotide polymorphism (rs2275
revealed after 2 hours of LPS stimulation of

Activated monocytes (~ 300 samples)

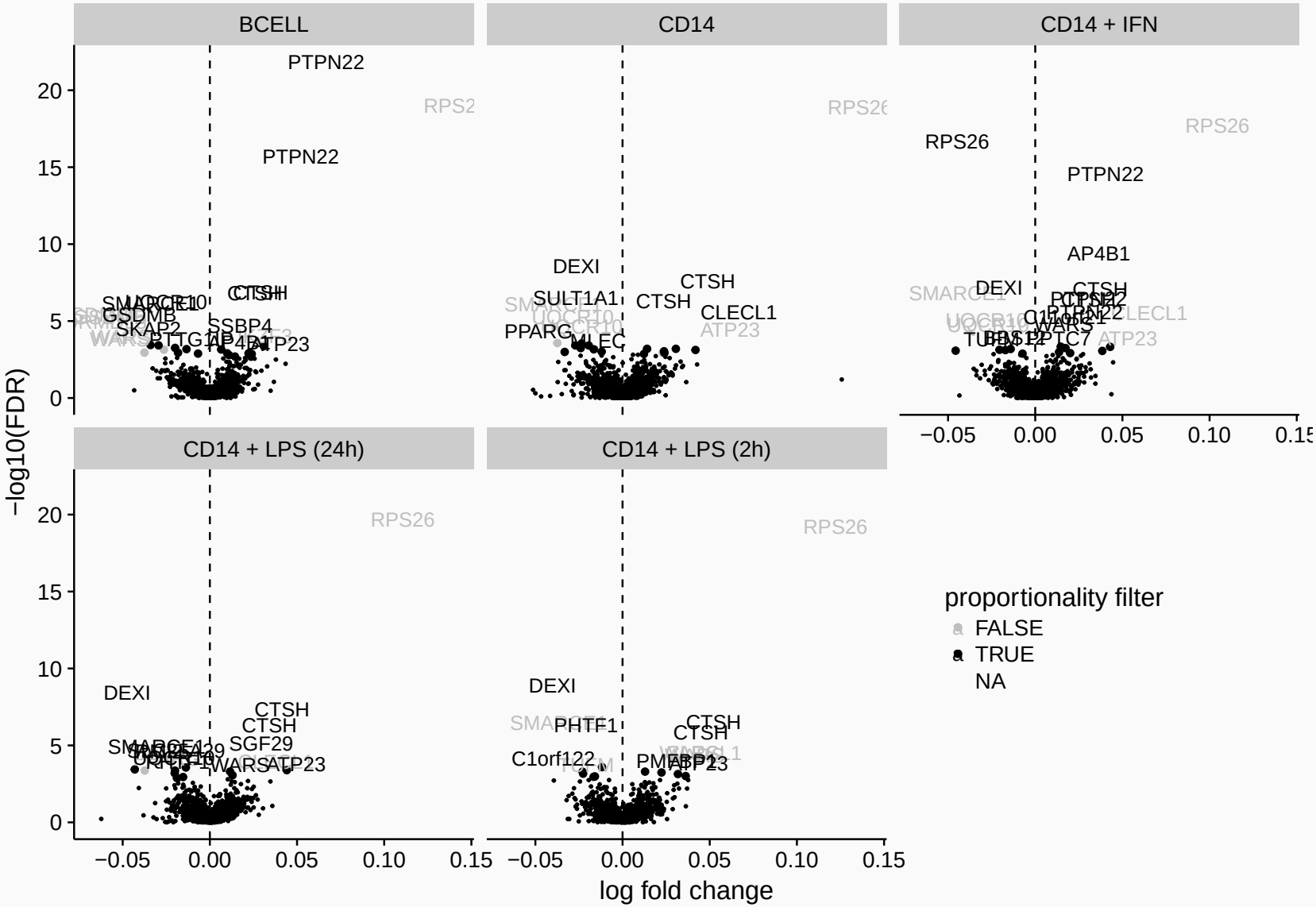
=

?

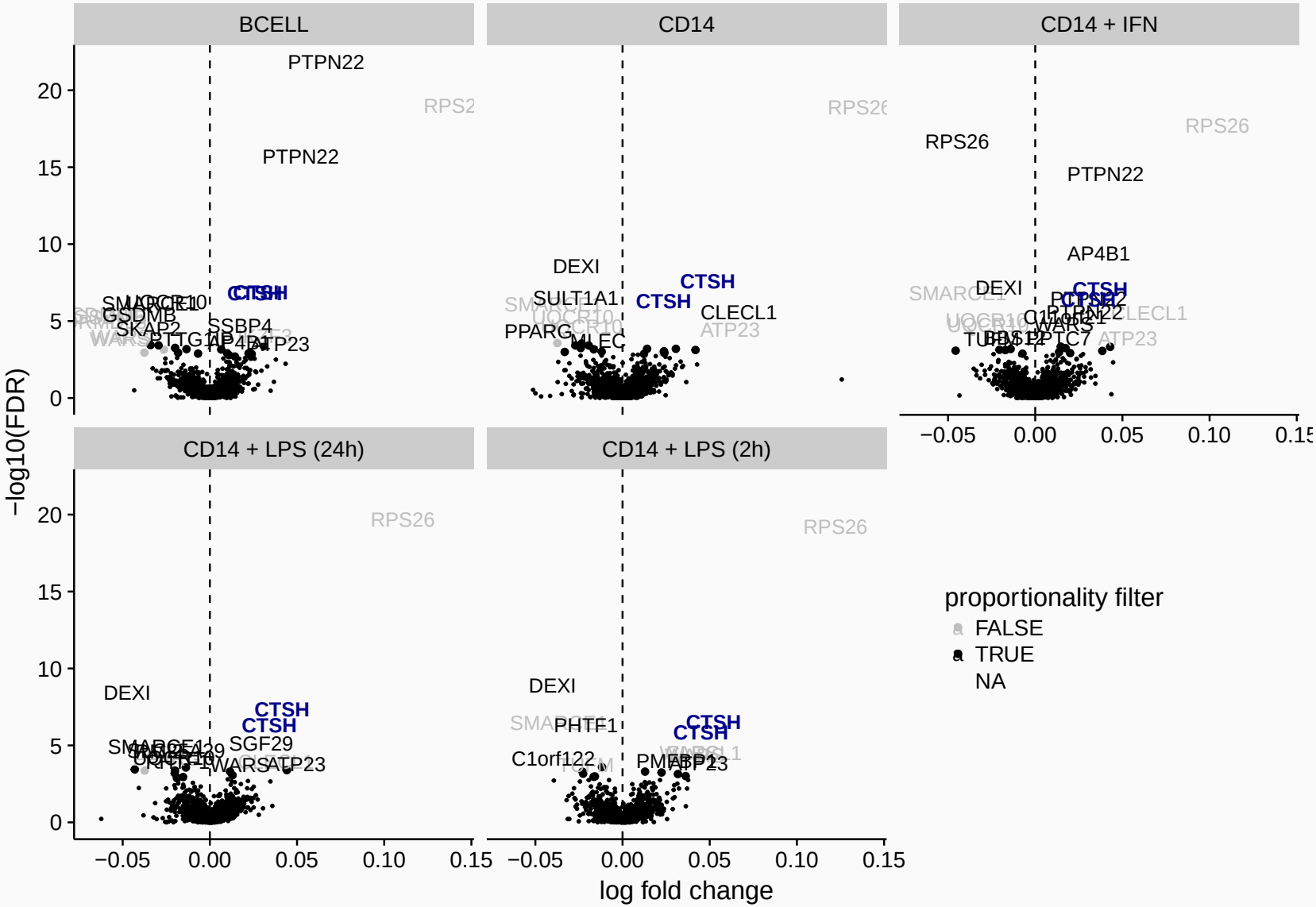
candidate causal autoimmune genes

Guo et al., Hum Mol Genet, *2015*

TWAS identified 88 T1D candidate genes, 61 after filtering



TWAS identified 88 T1D candidate genes, 61 after filtering



Summary of gene causal identification methods

Colocalisation

are GWAS and eQTL signals “the same”?

Close to sufficient for causality

very stringent hypothesis, possibility of false negatives

TWAS

what genes are (predicted to be) differentially expressed in patients? appear to be lots!

Possibly necessary, **NOT** sufficient evidence of causality

PCHi-C methods

does a GWAS variant lie in a region which contacts a gene promoter?

useful circumstantial evidence; **NOT** causality

Outline

- Inferring disease relevant cells
- Fine mapping causal genetic variants
- Causal genes and cells
- 3D genome regulation
- Colocalisation of disease and gene expression variants
- Transcriptome-wide association study

Remaining challenges and opportunities

Cell diversity

- Amazing diversity in gene expression, chromatin binding and chromatin conformation - we only study common, accessible cells!
- eQTL datasets need to be larger, more diverse
- PCHi-C *already* covers greater diversity than eQTL

Evidence synthesis

- Different methods, different assumptions $\rightarrow$ different genes

Wider adoption of methods

- Datasets need to be more available
- Methods need to be easier to apply