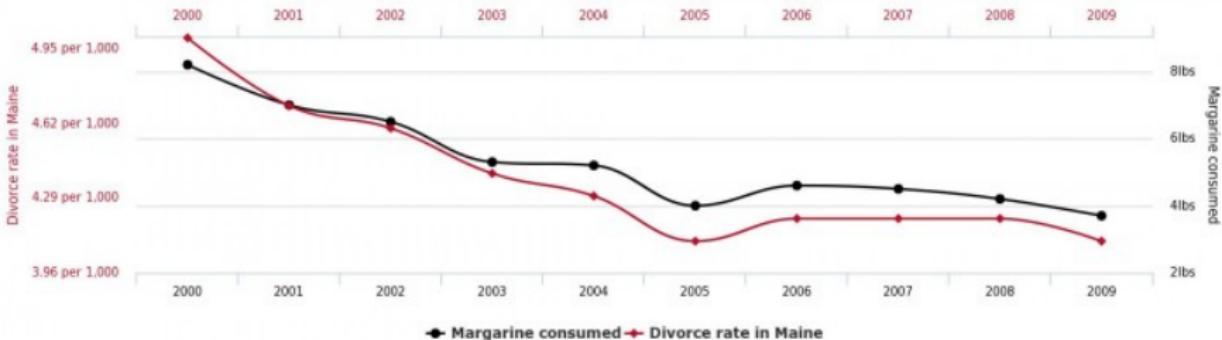


# GWAS: from association to causality

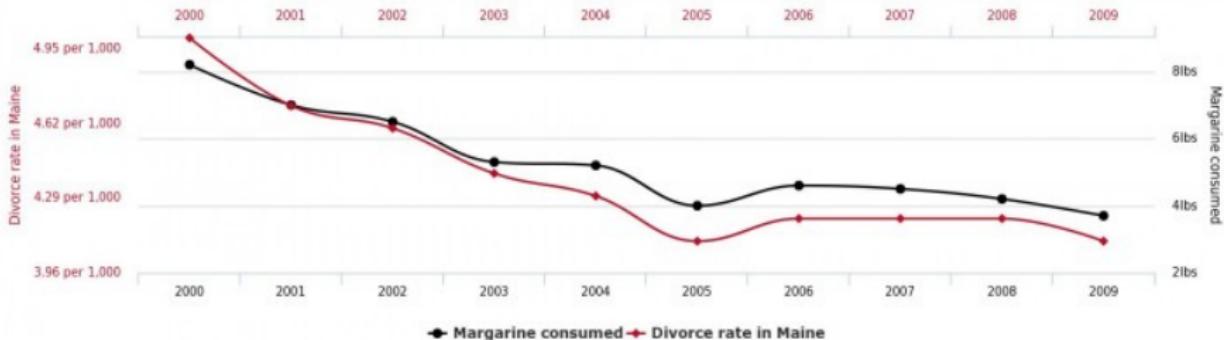
Chris Wallace

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

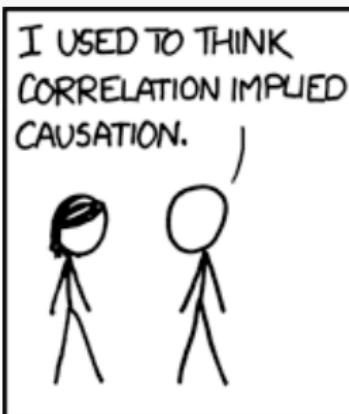
**Divorce rate in Maine**  
correlates with  
**Per capita consumption of margarine**



**Divorce rate in Maine**  
correlates with  
**Per capita consumption of margarine**



tylervigen.com



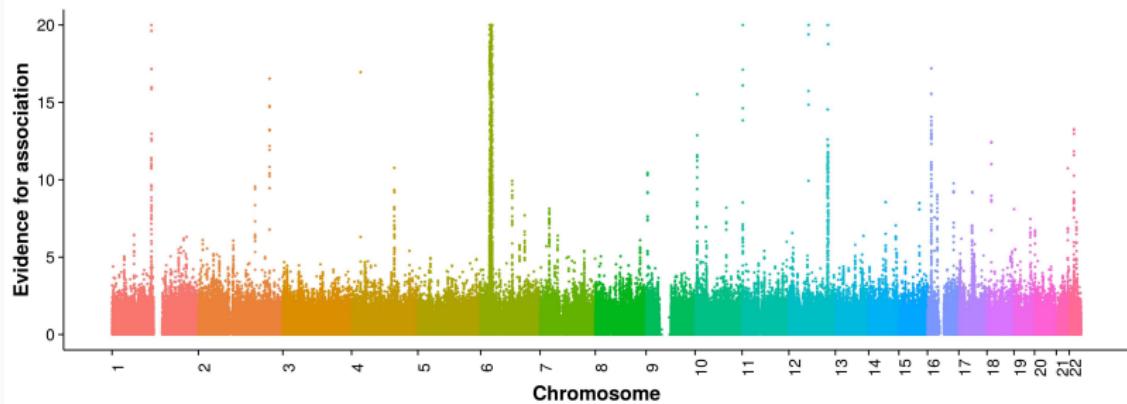
# Outline

- 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

---

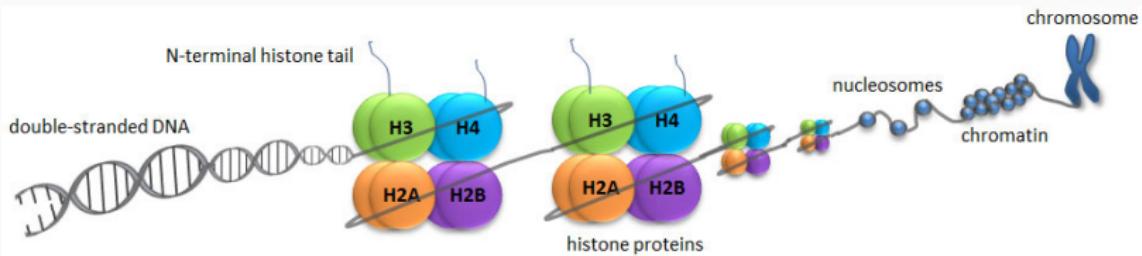
# Association does not identify the causal variant



# 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

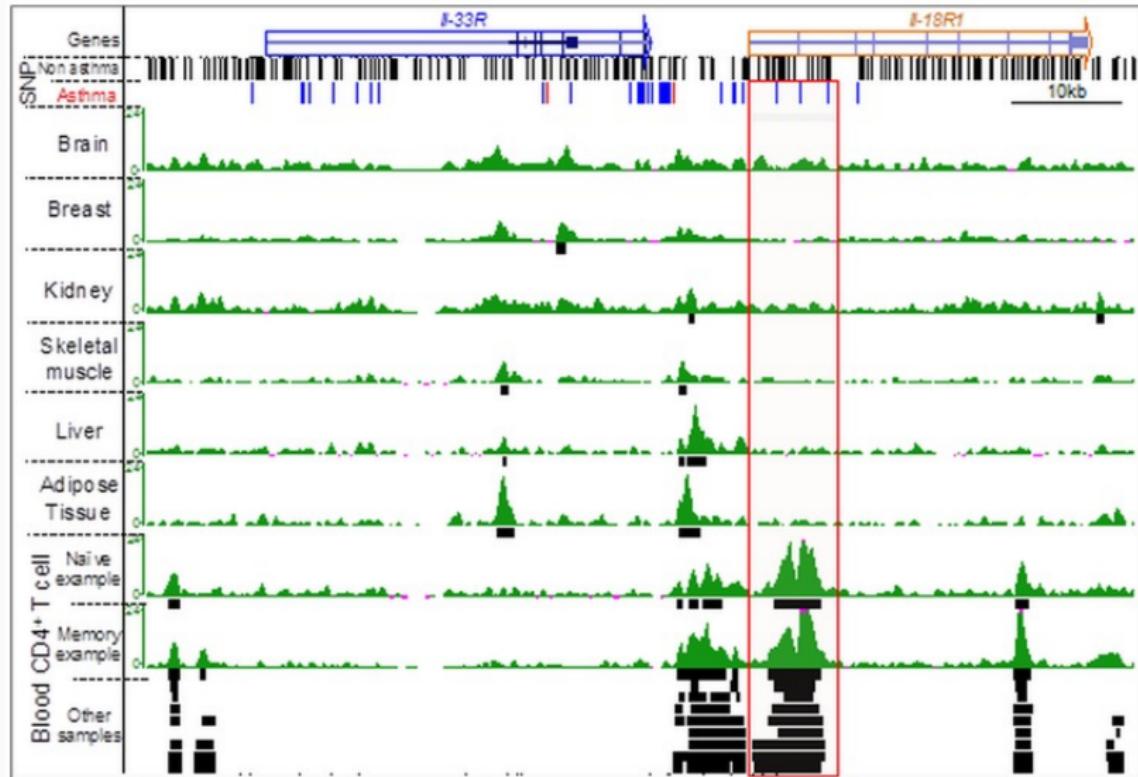


Resources:

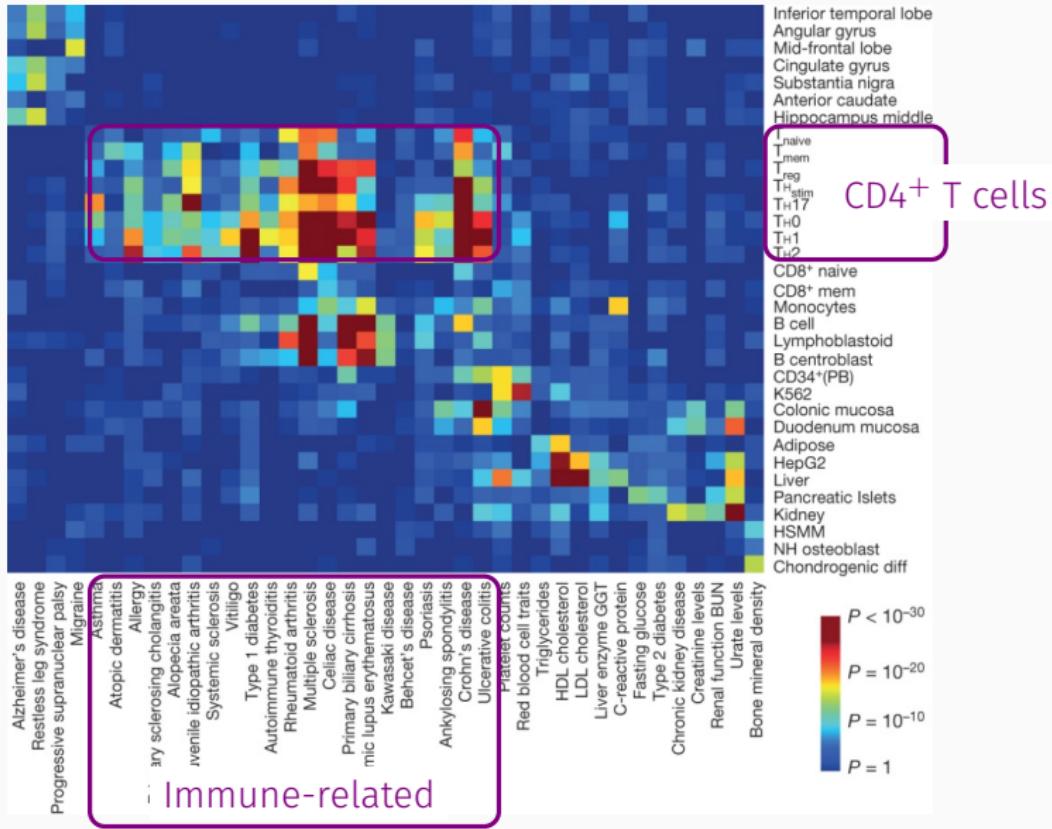
ENCODE <https://genome.ucsc.edu/encode>

BLUEPRINT <http://www.blueprint-epigenome.eu>

# We can learn something from location of associated variants



# Enrichment of enhancer marks links cells to disease

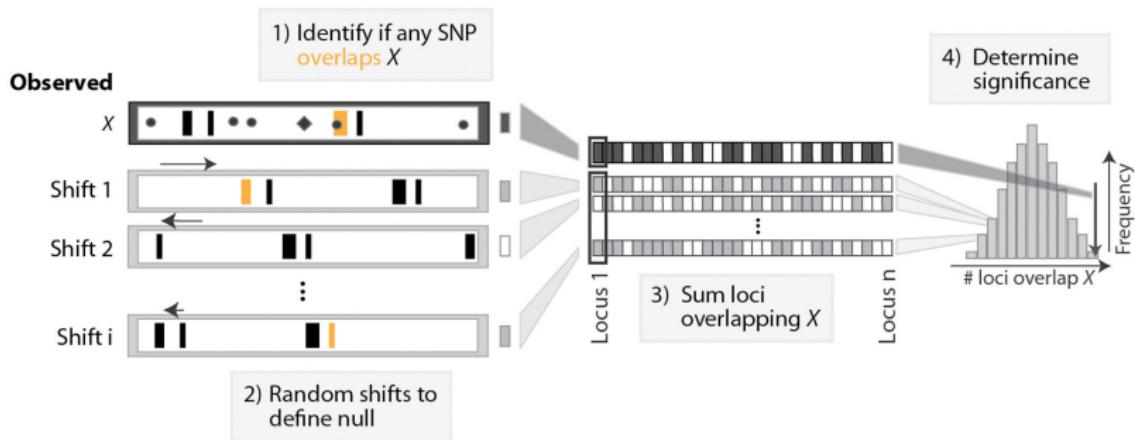


# LD (again)

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.



## 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?

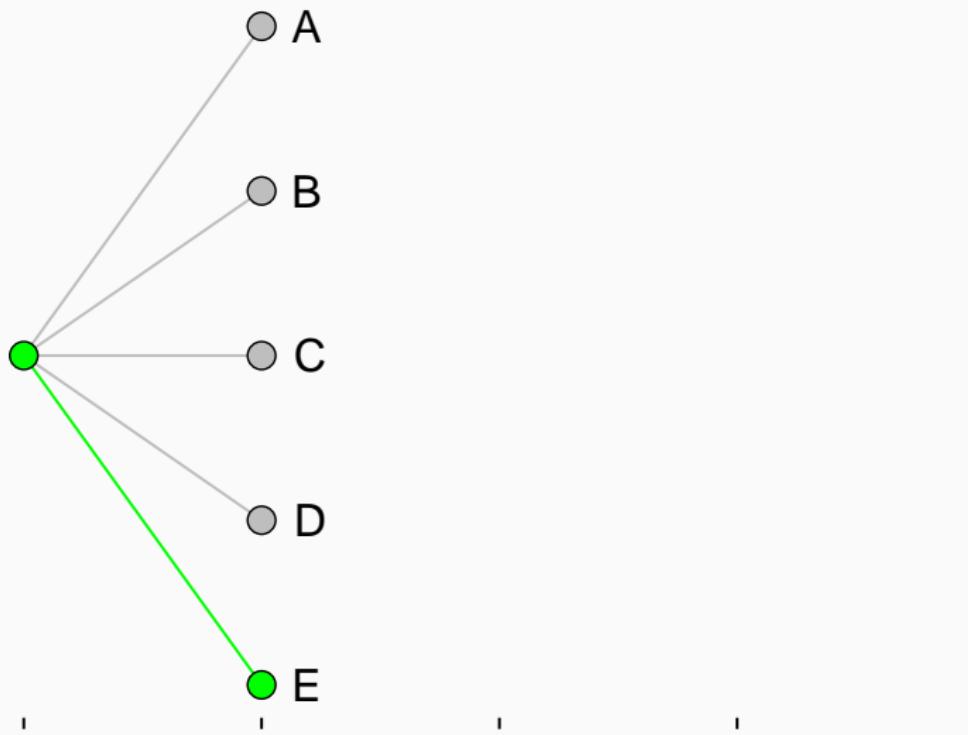
1: start with no SNPs



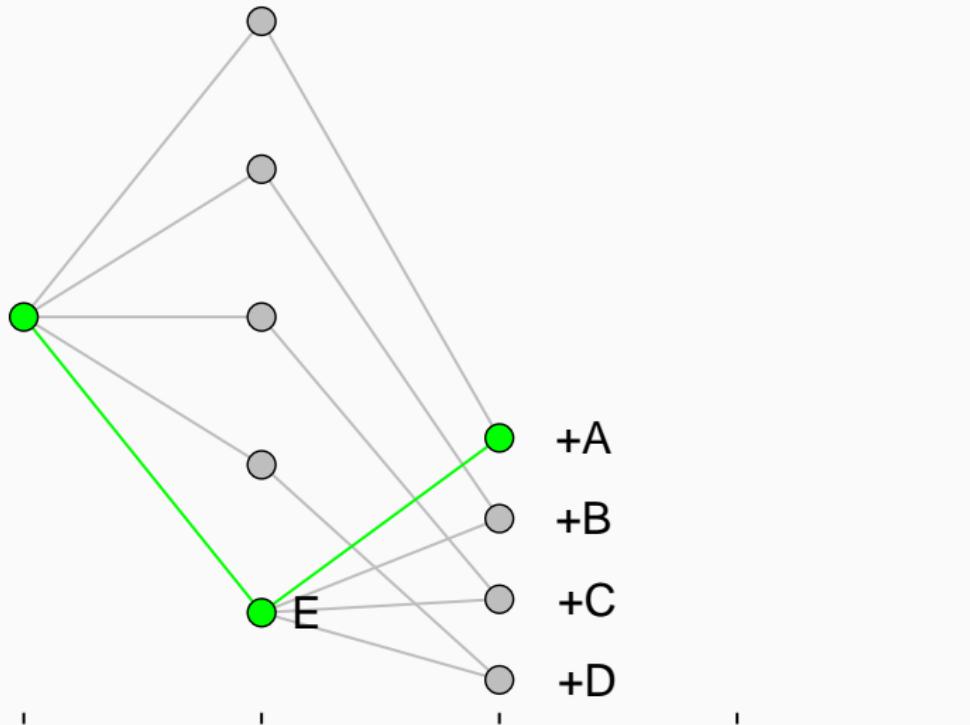
0 SNPs



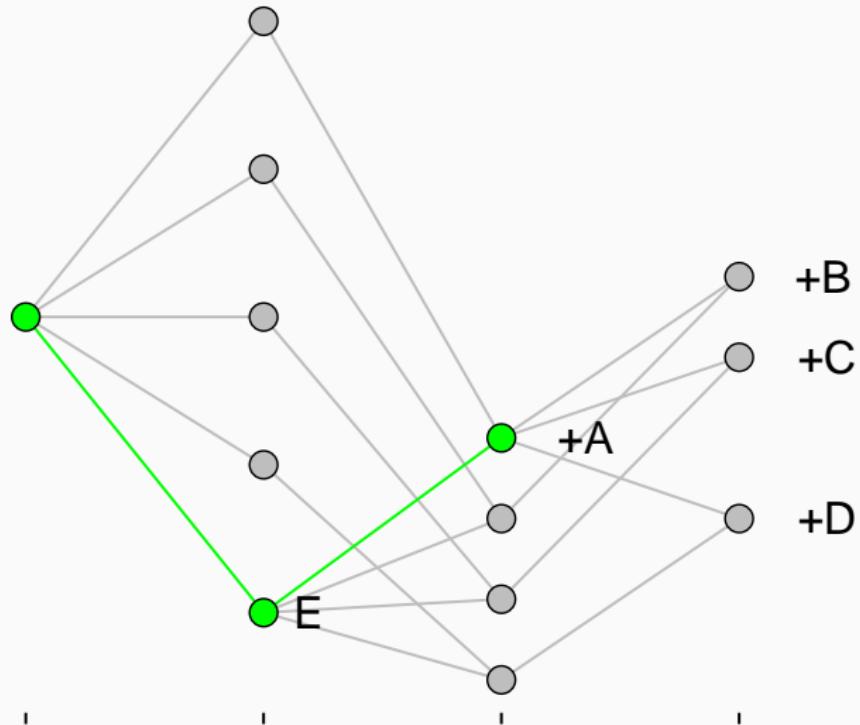
## 2: best model assuming single causal variant



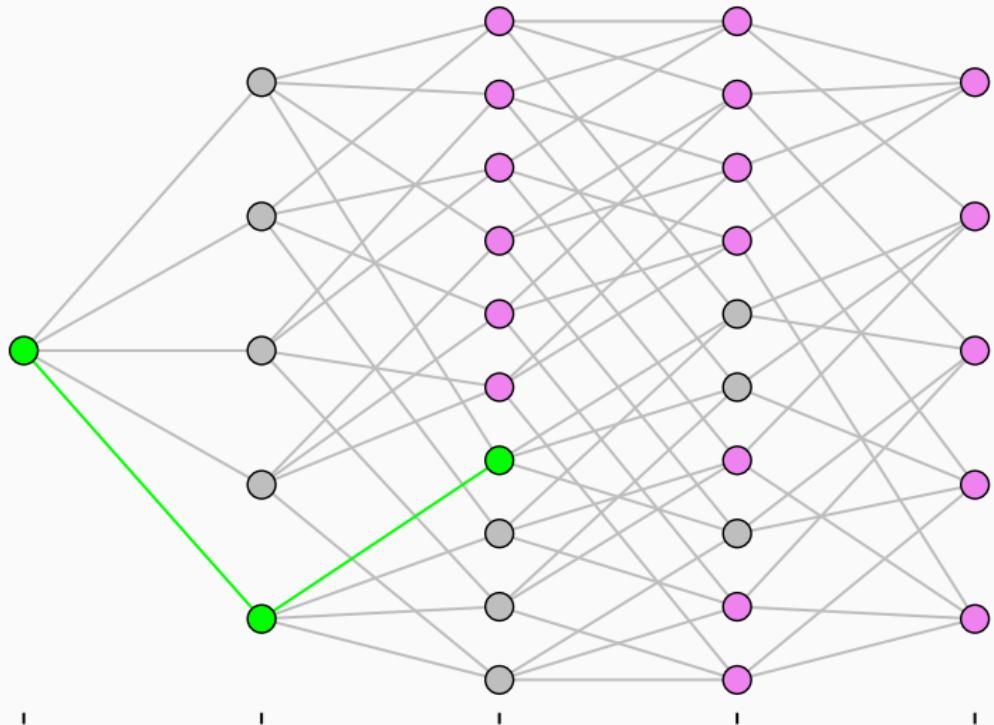
### 3: step forward into two SNP models



#### 4: consider three SNP models



We only explored a subset of models



## Alternative: regularized regression

### OLS regression

$$\text{Objective function: } \sum_i (y_i - \beta^T x_i)^2$$

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  $\beta$  “shrunk” to 0.

$$\text{Objective function: } \sum_i (y_i - \beta^T x_i)^2 + \lambda \sum_j |\beta_j|$$

$\lambda$  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	$1000 * 999 * 998 / (2 * 3) = 166,167,000$
$\vdots$	

## Joint modelling under single causal variable assumption

Consider  $n$  SNPs.

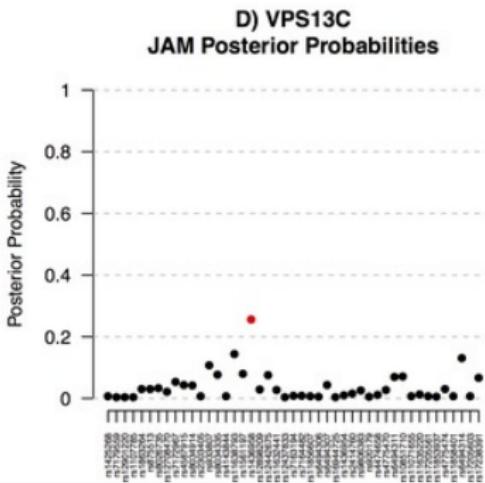
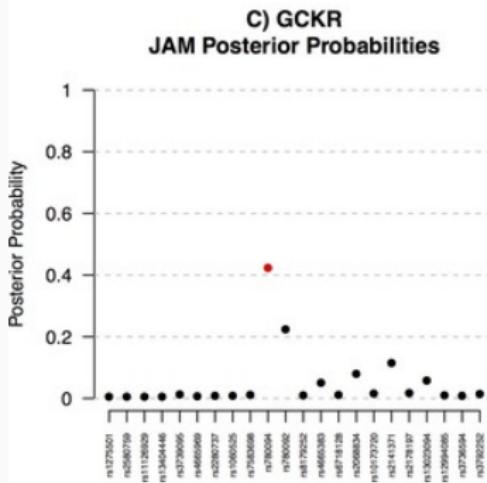
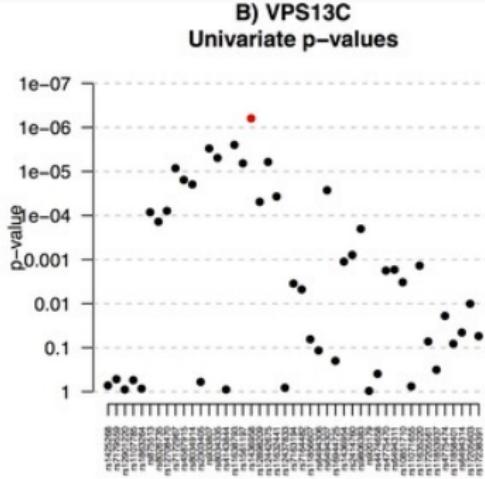
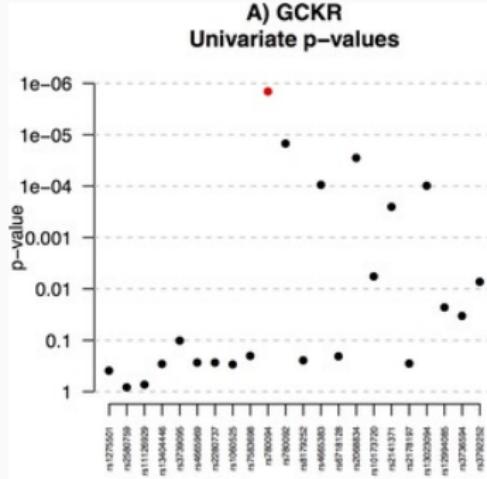
Assume all 1 SNP models equally likely with Probability  $\pi_1$  (or not - if you have external information)

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

$P(\text{SNP } i \text{ causal}) =$

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

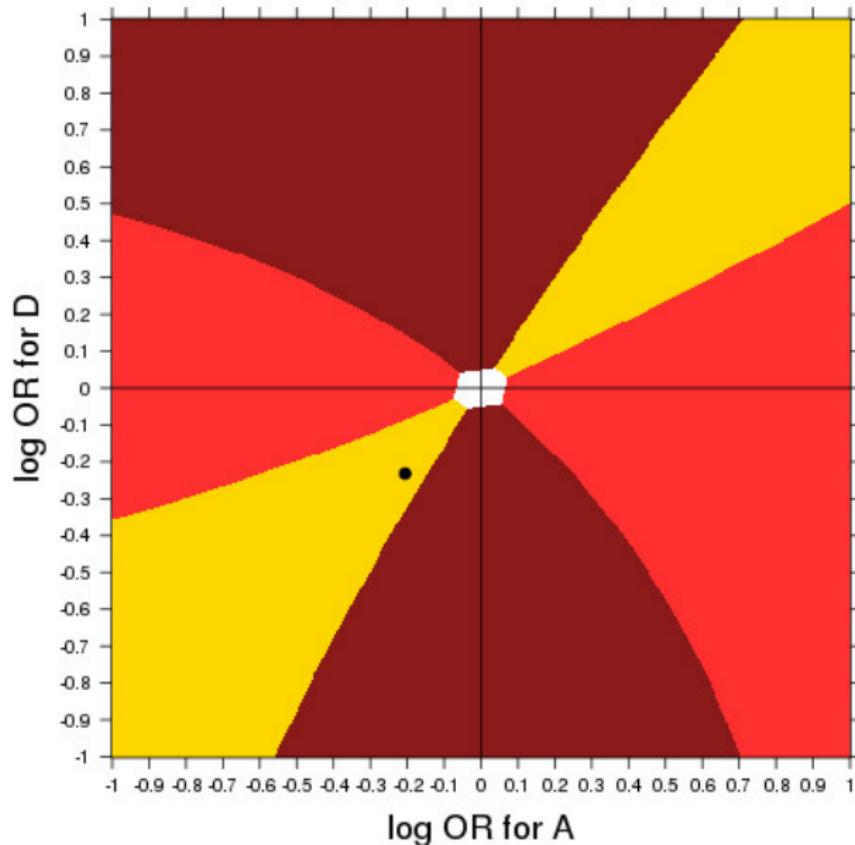
*Can calculate these probabilities via Bayes factors. Wakefield (2009)<sup>3</sup> gives methods to calculate approximate Bayes factors from GWAS p values*



## The first SNP might not be (close to) causal

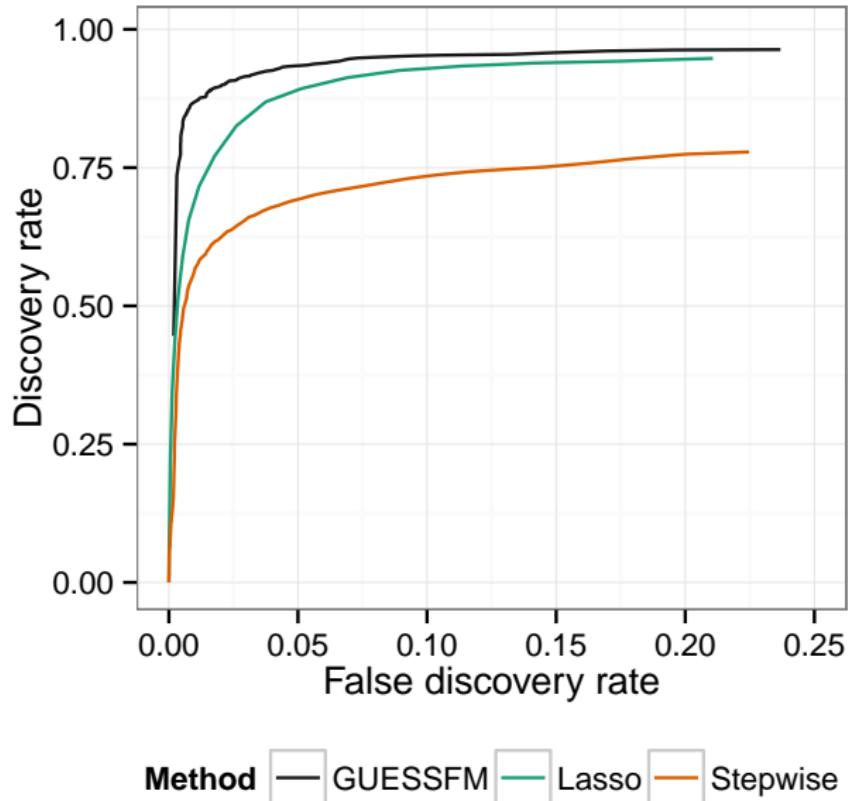
	rs2104286	A: rs12722496	D: rs56382813
p value	7.61E-23	1.16E-05	6.45E-18
$r^2$			
rs2104286	1.00	0.33	0.31
A: rs12722496	0.33	1.00	0.08
D: rs56382813	0.31	0.08	1.00

The first SNP might not be (close to) causal

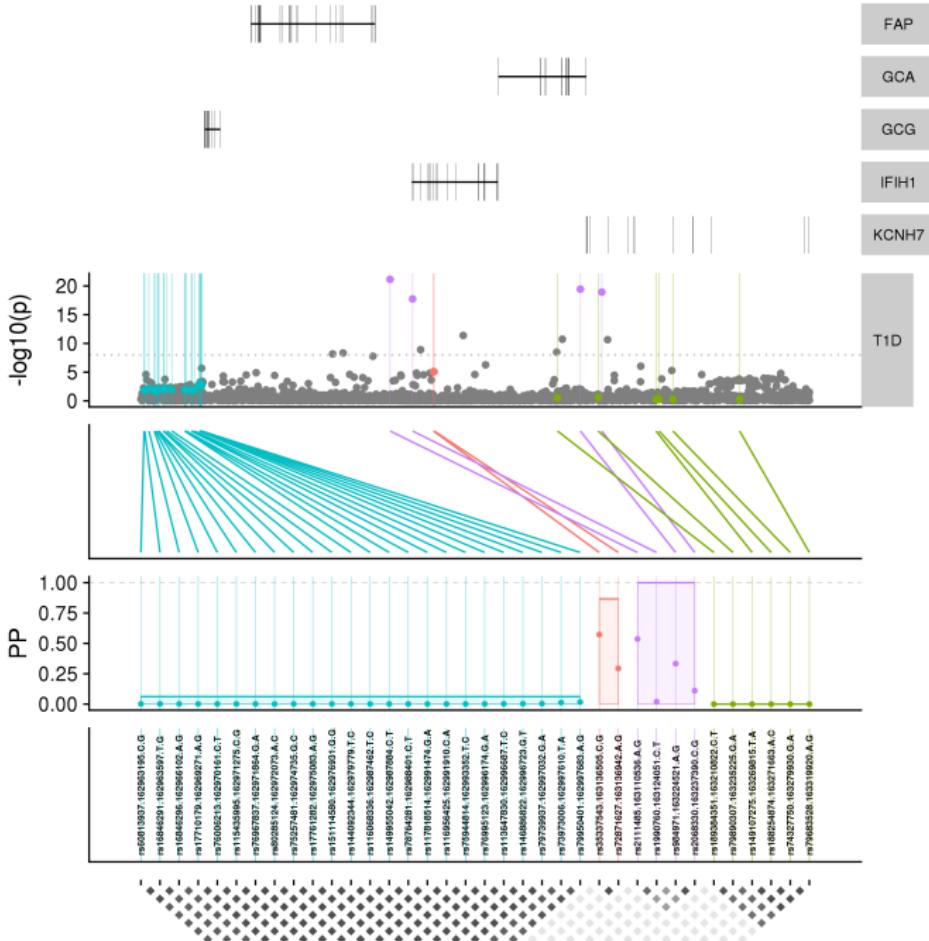


## Alternative (3): joint modelling with ‘stochastic search’

## Better recovery of “causal” variants in simulated data



2q-162960873-163361685



## 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: Bradford Hill and genetics

Strength

Temporality

Coherence

Consistency

Biological gradient

Experiment

Specificity

Plausibility

Analogy

BIOLOGY IS LARGELY SOLVED.  
DNA IS THE SOURCE CODE  
FOR OUR BODIES. NOW THAT  
GENE SEQUENCING IS EASY,  
WE JUST HAVE TO READ IT.

IT'S NOT JUST "SOURCE  
CODE." THERE'S A TON  
OF FEEDBACK AND  
EXTERNAL PROCESSING.



BUT EVEN IF IT WERE, DNA IS THE  
RESULT OF THE MOST AGGRESSIVE  
OPTIMIZATION PROCESS IN THE  
UNIVERSE, RUNNING IN PARALLEL  
AT EVERY LEVEL, IN EVERY LIVING  
THING, FOR FOUR BILLION YEARS.

IT'S STILL JUST CODE.



OK, TRY OPENING GOOGLE.COM  
AND CLICKING "VIEW SOURCE."

OK, I... OH MY GOD.

THAT'S JUST A FEW YEARS OF  
OPTIMIZATION BY GOOGLE DEV'S.  
DNA IS THOUSANDS OF TIMES  
LONGER AND WAY, WAY WORSE.

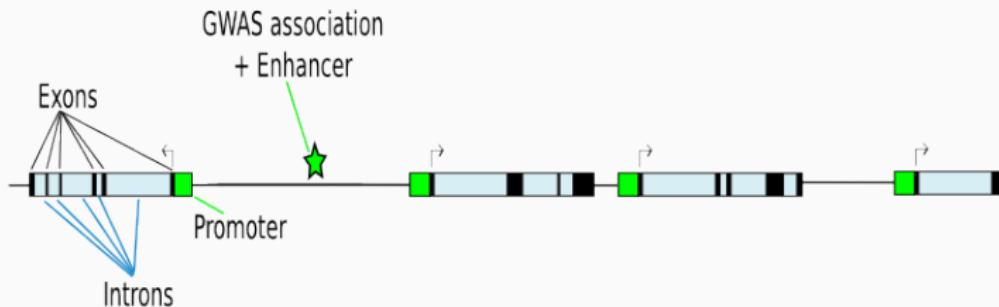
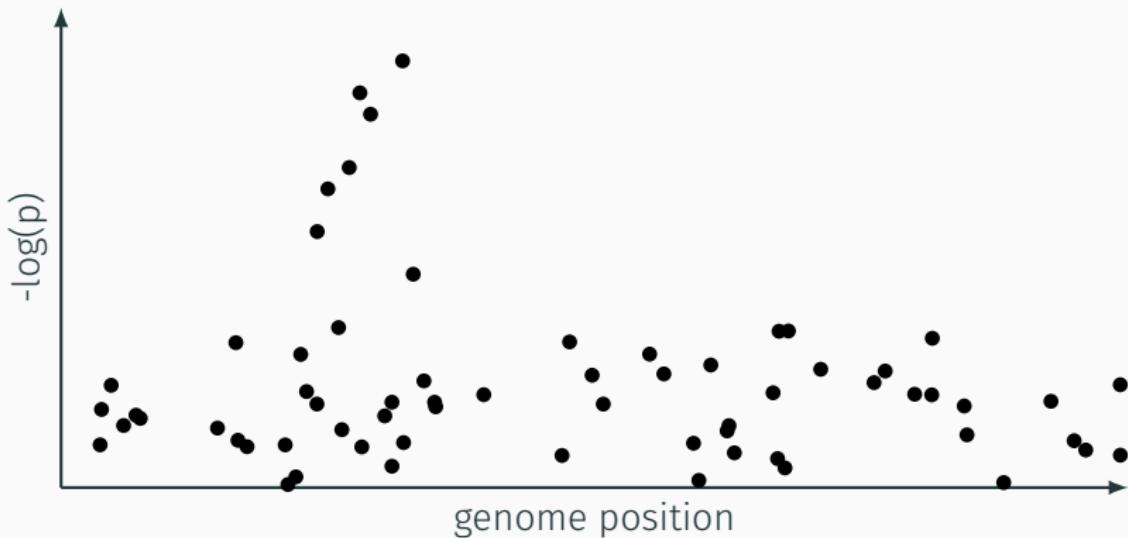
WOW, BIOLOGY  
IS IMPOSSIBLE.



## Causal genes and cells

---

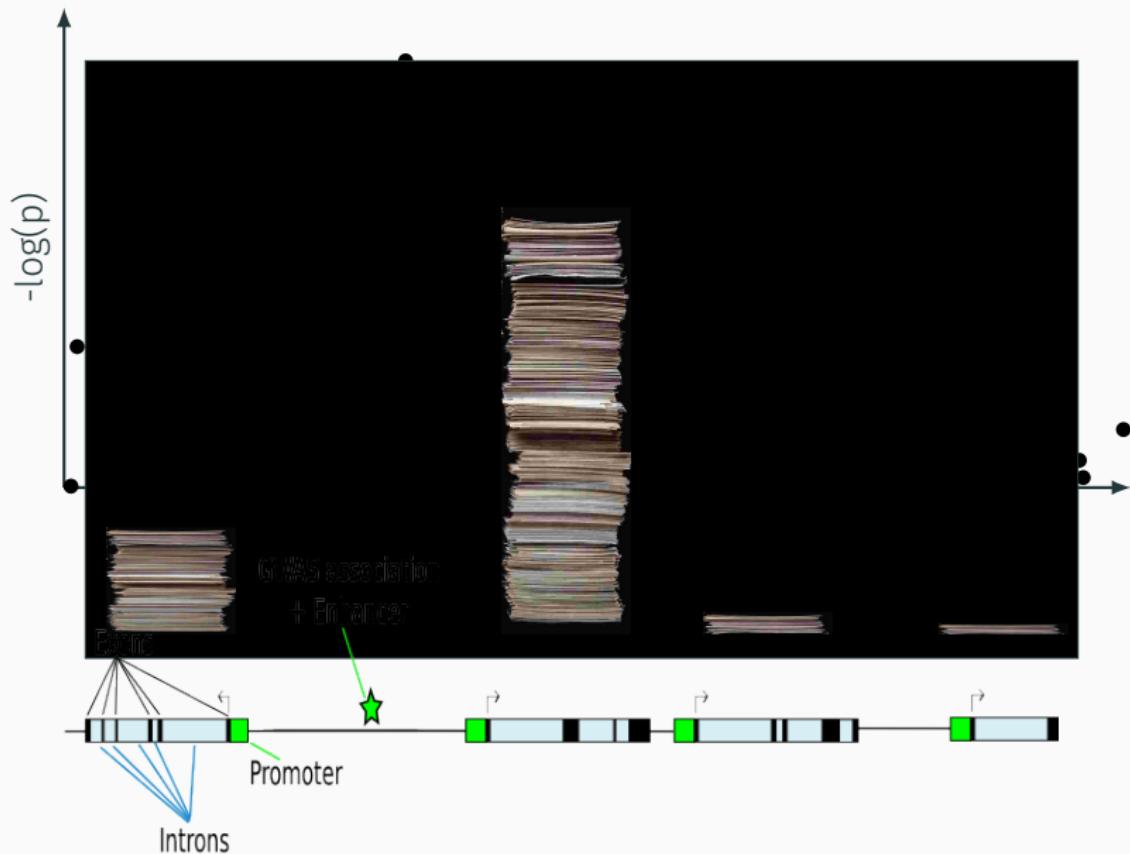
# Which *gene* alters disease risk?



Which *gene* alters disease risk?



# Which *gene* alters disease risk?



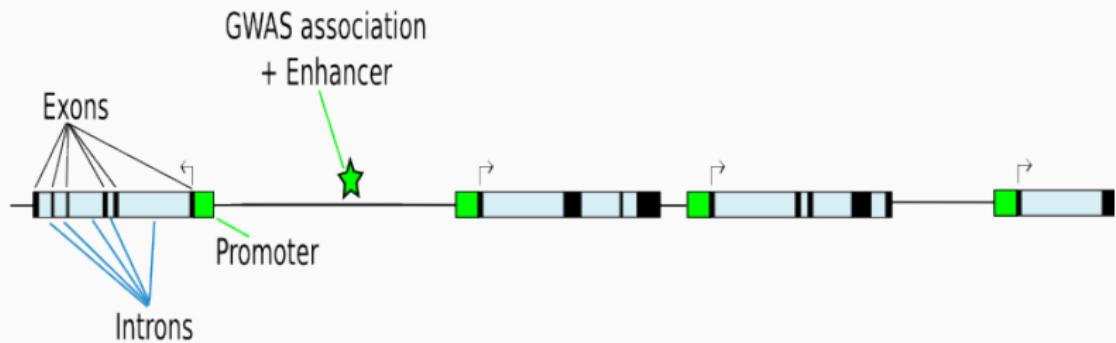
# Which *gene* alters disease risk?



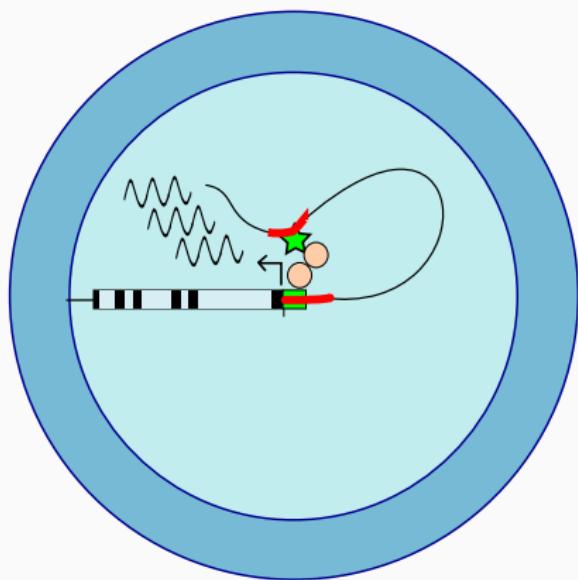
## 3D genome regulation

---

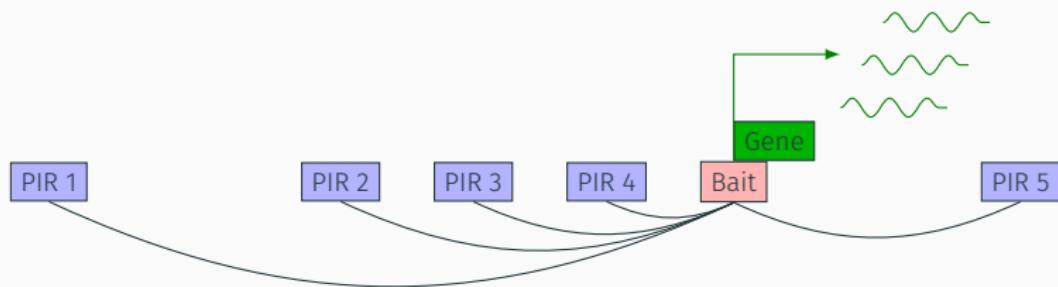
# 3D folding of DNA in nucleus connects enhancers to promoters



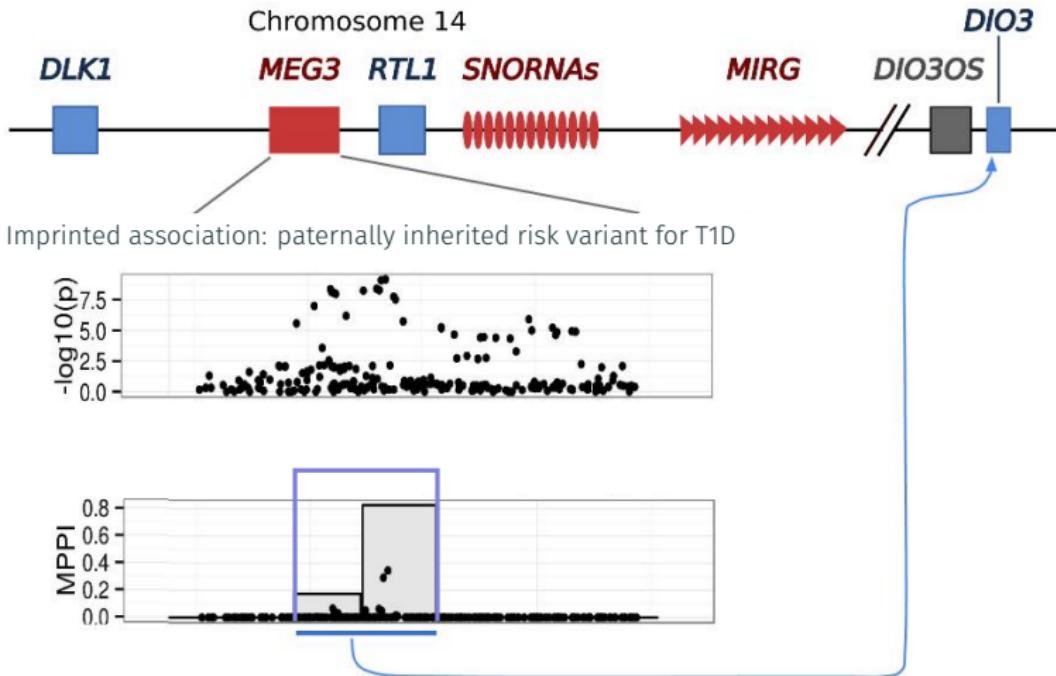
# 3D folding of DNA in nucleus connects enhancers to promoters



# Promoter capture Hi-C



# Integrate GWAS + PChi-C to prioritise candidate disease genes



## Colocalisation of disease and gene expression variants

---

# Integrate GWAS and eQTL datasets via colocalisation

0.5-8M variants

genotypes

> 10k genes

expression

100s individuals

summaries

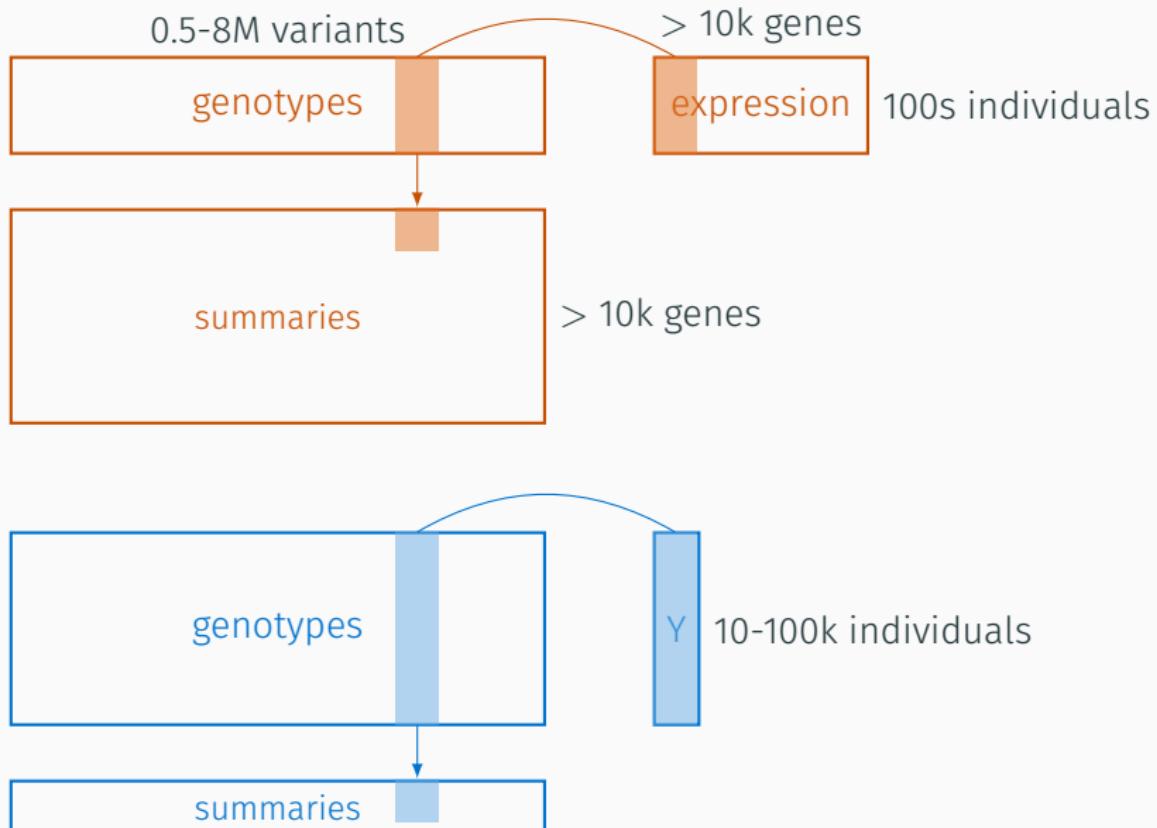
> 10k genes

genotypes

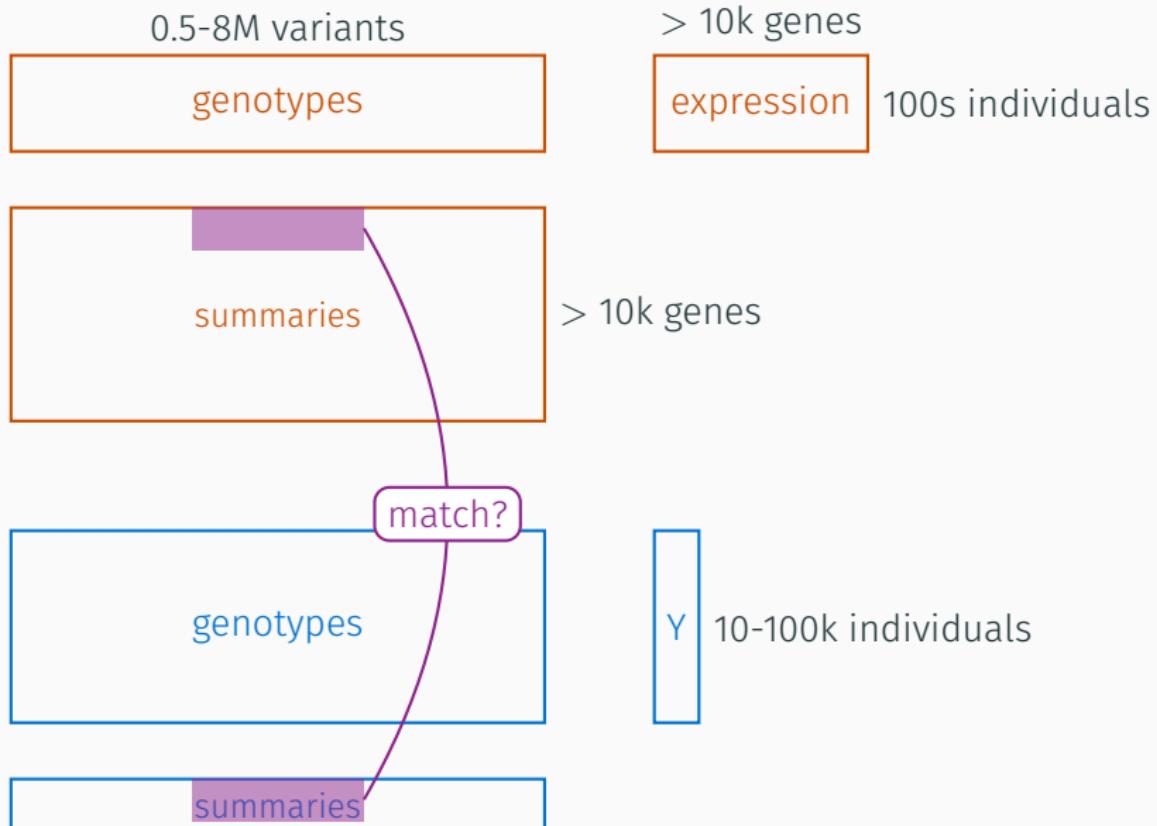
Y 10-100k individuals

summaries

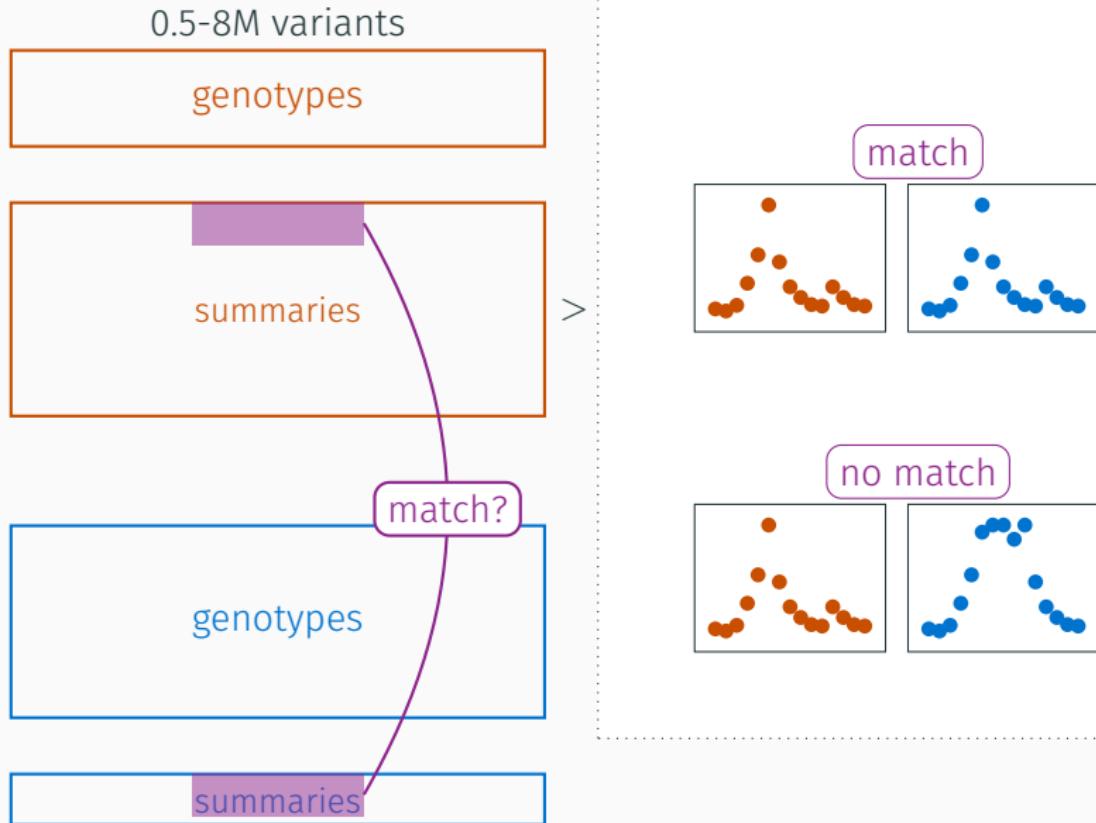
# Integrate GWAS and eQTL datasets via colocalisation



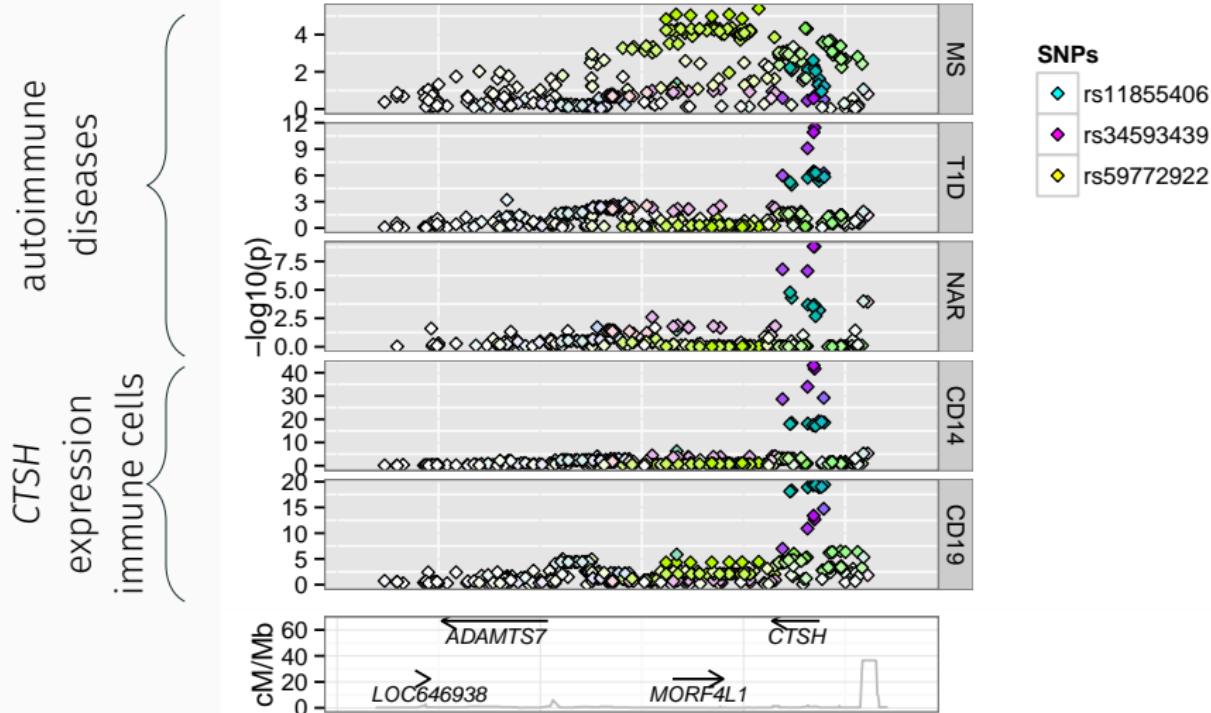
# Integrate GWAS and eQTL datasets via colocalisation



# Integrate GWAS and eQTL datasets via colocalisation



# Colocalisation points to *CTSH* in CD14<sup>+</sup> monocytes



# Colocalisation analysis 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<sup>1</sup>, Seiko Makino<sup>1</sup>, Jayachandran Radhakrishnan<sup>1</sup>, Katharine Plant<sup>1</sup>, Stephen Leslie<sup>2</sup>, Alexander Dilthey<sup>3</sup>, Peter Ellis<sup>4</sup>, Cordelia Langford<sup>4</sup>, Fredrik O Vannberg<sup>1,5</sup> & Julian C Knight<sup>1</sup>

Trans-acting genetic variants impacting gene expression in primary immune cells expression quantitative trait cells, respectively. Add disease locus, with P/QC supported by evidence of cell type-specific associations with all pathogens. We also find AGAP and ARHGAP24 i cell populations identify disease susceptibility.

Defining the genetic determinants of gene expression regulation. This is particularly i variants underlying other wide association studies (I linkage disequilibrium) in the context of relevant cell state<sup>2–5</sup>. This context-dependent

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

Benjamin P. Fairfax,\* Peter Hünig, Seiko Makino, Vivek Narayanan, Daniel Weng, Evelyn Lai, Luke Jordan, Katherine Plant, Robert Andrews, Chris McGee, Julian C. Knight\*

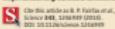
**Introduction:** Many genetic variants associated with common disease susceptibility occur close to regulatory elements. Defining the function of these variants requires identifying the specific functional variants and the specific genes that they regulate remains challenging and in many cases unassessable. We hypothesized that a significant proportion of variants, including those implicated in disease, show effects in a context-specific manner and therefore can be identifiable upon triggering of relevant responses.

**Method:** We mapped individual 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 prepared primary CD14+ monocytes from healthy volunteers exposed to the innate inducer lipopolysaccharide (LPS) or different degrees (2 or 24 hours) of liposomal doxorubicin (LDO). eQTL mapping was performed on a genome-wide basis with an additive linear model. A subset of 220 individuals with expression data available for all experimental conditions enabled cross-treatment comparisons.

**Results:** Stimulation with LPS or LDO resulted in profound changes 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 interacted established canonical pathways of monocyte signaling and included key nodal genes and effector molecules. These eQTLs are typically most active at the transcription start site and, in some cases, showed reversal of

READ THE FULL ARTICLE ONLINE

<http://dx.doi.org/10.1126/science.12446>



FIGURES IN THE FULL ARTICLE

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

Fig. 2. Trans-eQTLs demonstrate context specificity and identify master regulatory elements after LPS.

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

Fig. 4. Cis regulation of IFN $\beta$  at rs1131494 has profound transcriptional consequences.

Fig. 5. Stimulus-specific eQTL and GRAS5 informative for disease risk.

SUPPLEMENTARY MATERIALS

Materials and Methods

Figs. S1 to S2

## Activated monocytes (~ 300 samples)

In contrast, 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 *ADM*, *IFNE*,

class with a  
capping  
IFN $\beta$  expression  
a virgin-vaccinated primate-specific eQTL  
measured after 2 hours of LPS stimulation of a



?

candidate causal  
autoimmune genes

# Colocalisation identified six candidate causal autoimmune disease genes

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

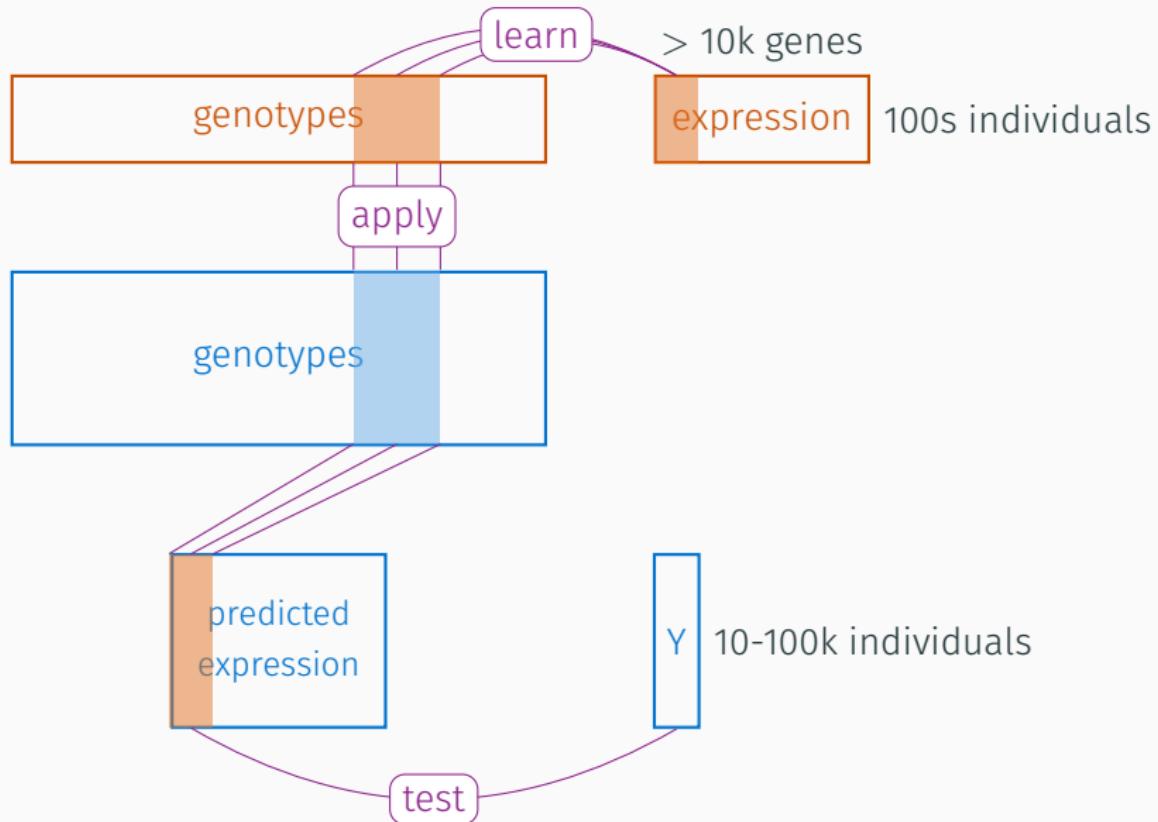
## Transcriptome-wide association study

---

# Integrate GWAS and eQTL datasets via TWAS

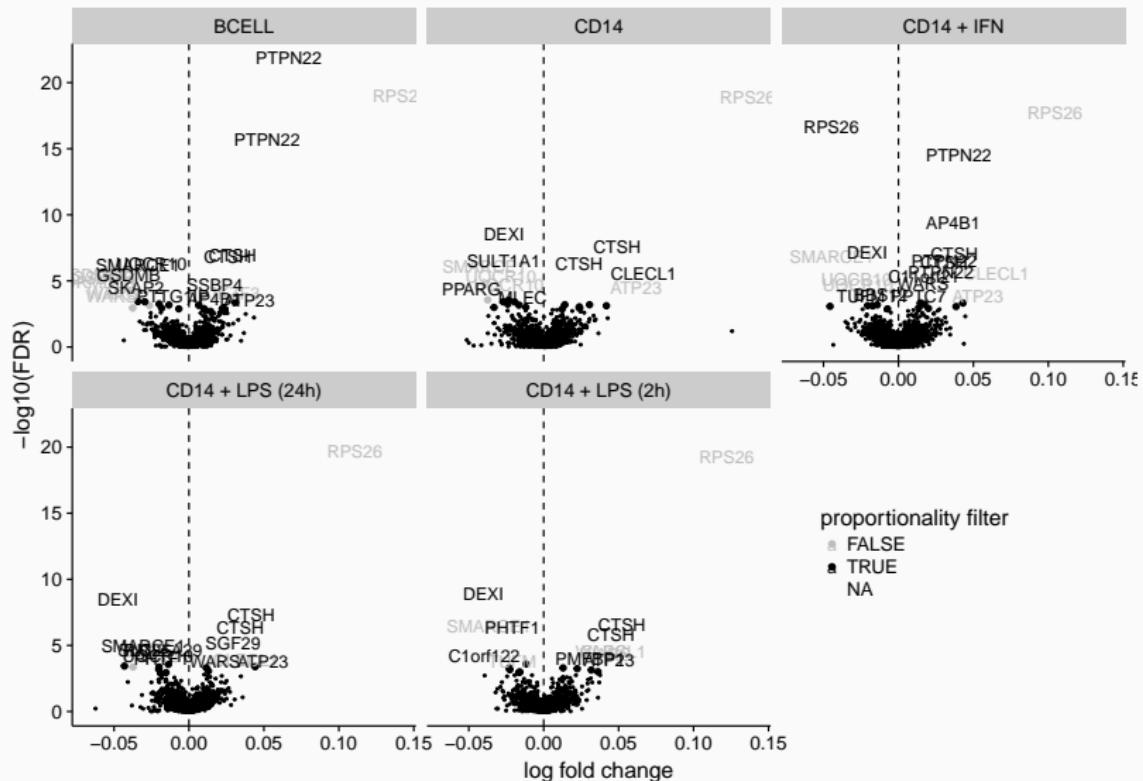


# Integrate GWAS and eQTL datasets via TWAS

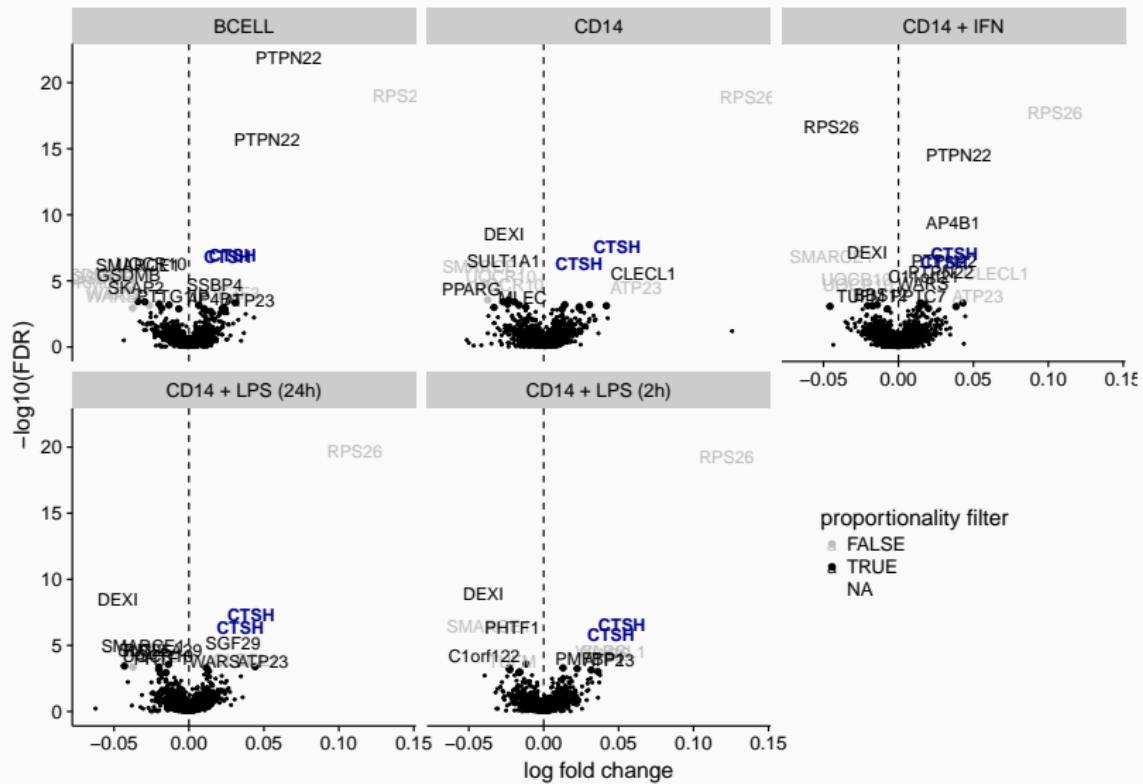




# TWAS identified 88 T1D candidate genes, 61 after filtering



# 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 → different genes

## Wider adoption of methods

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

