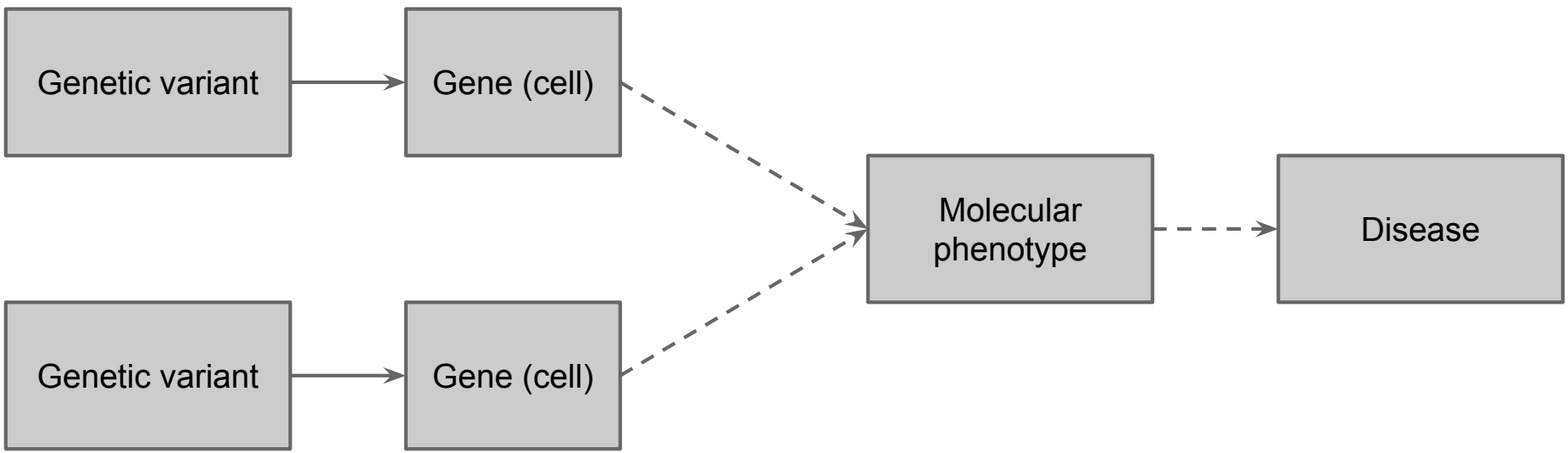
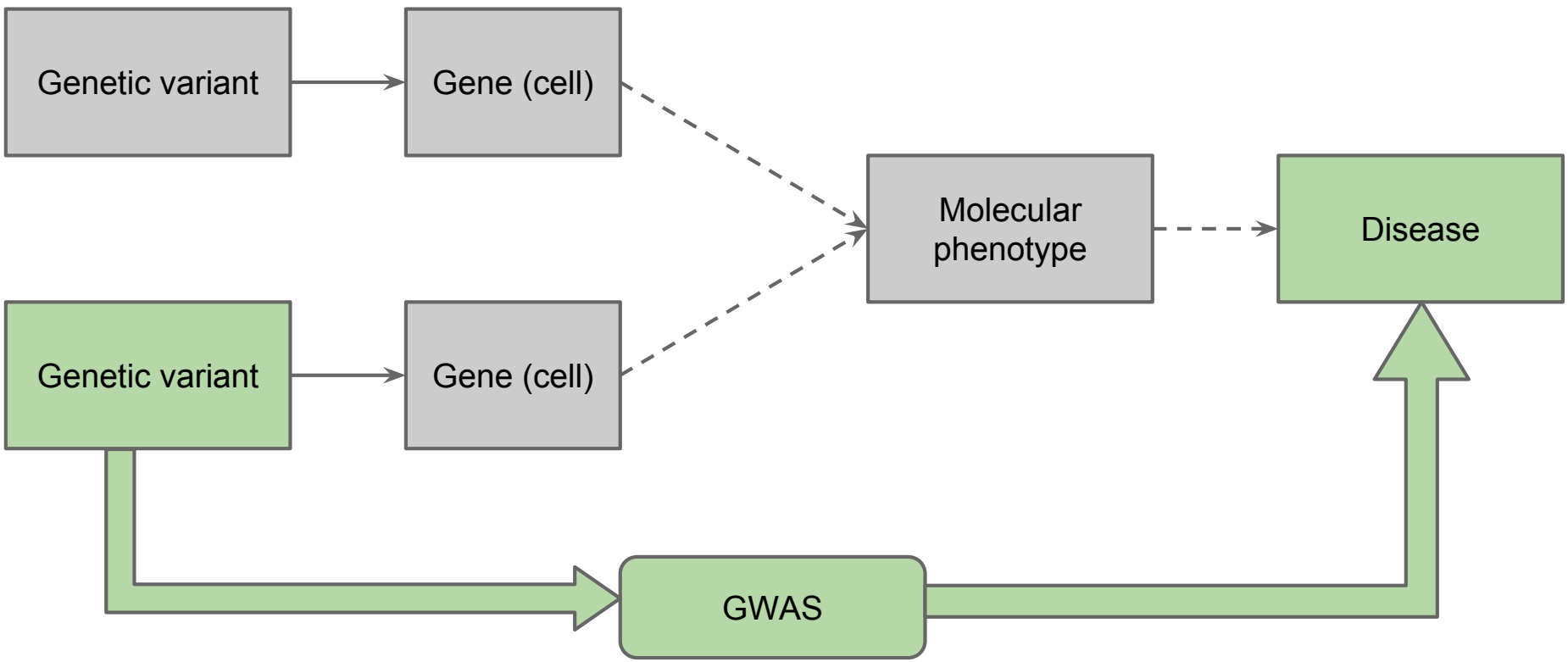
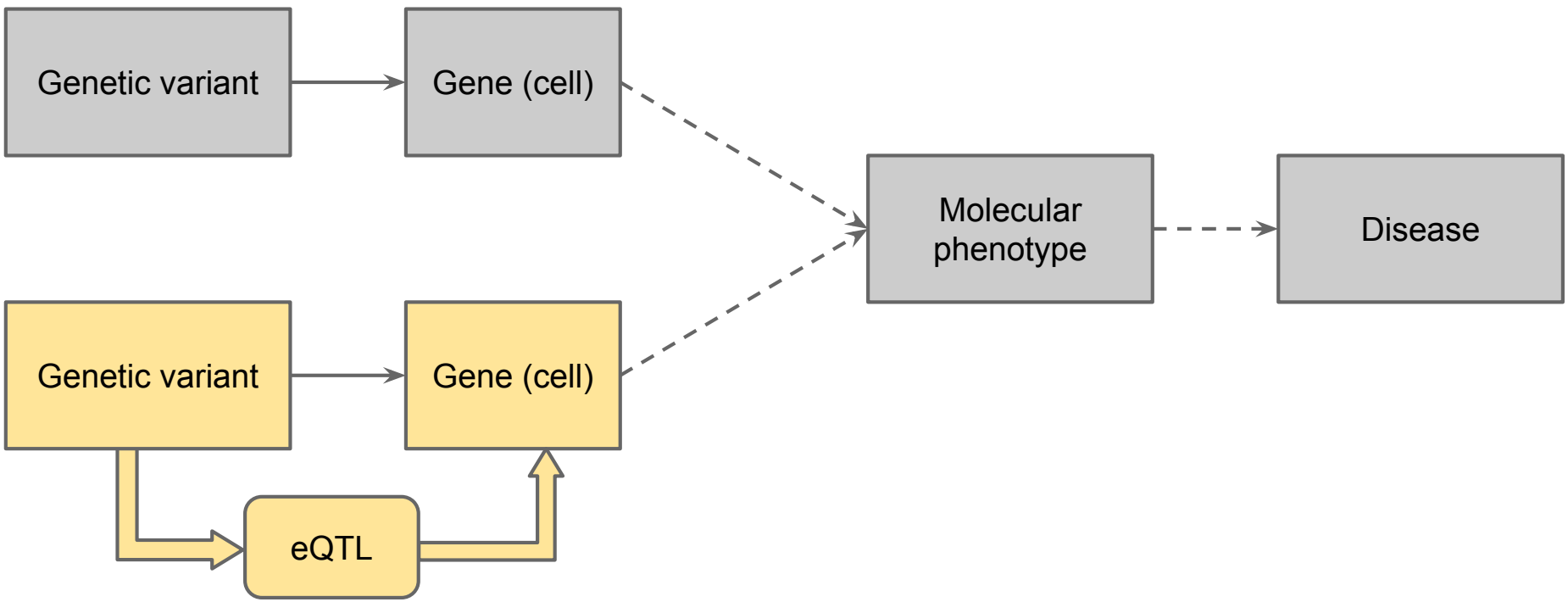
From genomewide association studies to causal variants and genes

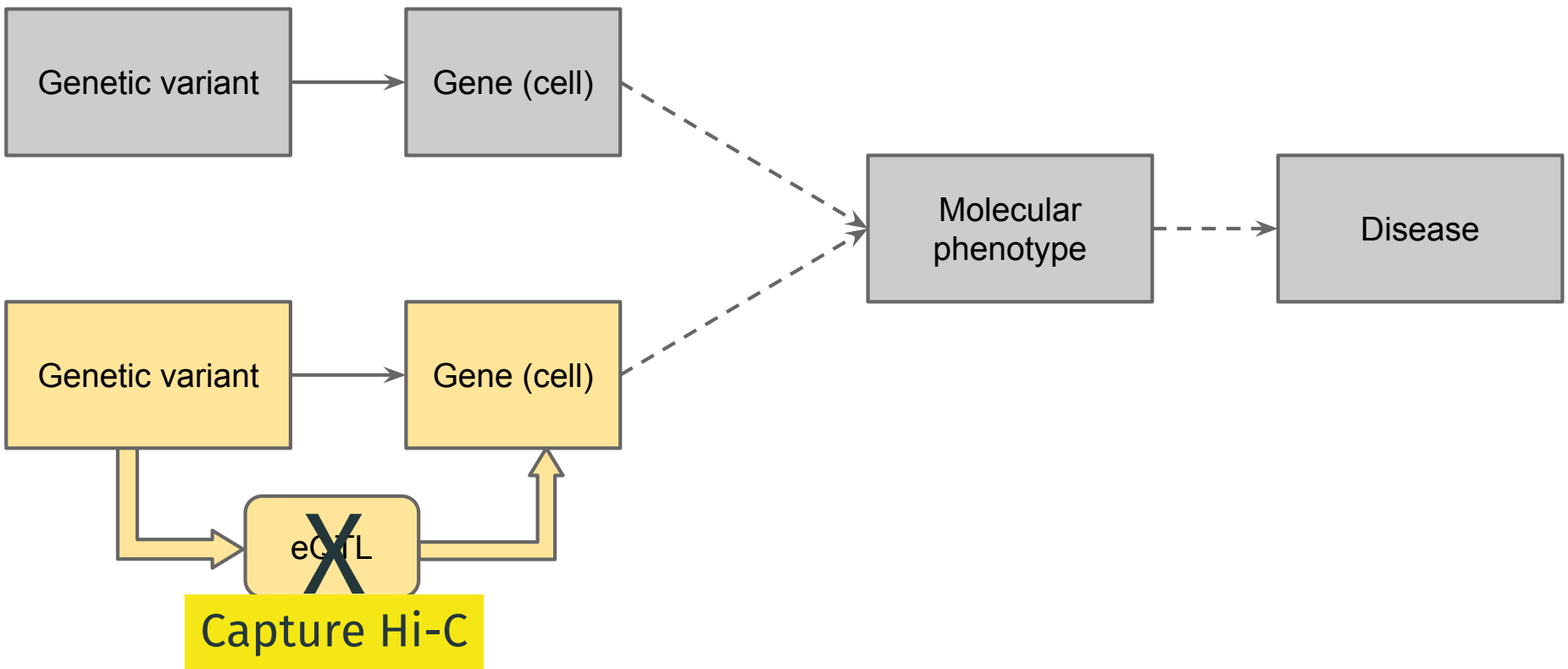
Chris Wallace

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


Genetic variant
Gene (cell)
Genetic variant
Gene (cell)
Molecular phenotype
Disease


Genetic variant
Gene (cell)
Molecular phenotype
Disease
Genetic variant
Gene (cell)
GWAS


Genetic variant
Gene (cell)
Molecular phenotype
Disease
Genetic variant
Gene (cell)
eQTL


Genetic variant
Gene (cell)
Molecular phenotype
Disease
Genetic variant
Gene (cell)
eQTL
Capture Hi-C

Outline

- Genomewide association studies: a brief introduction
- Inferring disease relevant cells
- Mapping causal variants
- Identifying causal genes from causal variants
- Summary

Genomewide association studies: a brief introduction

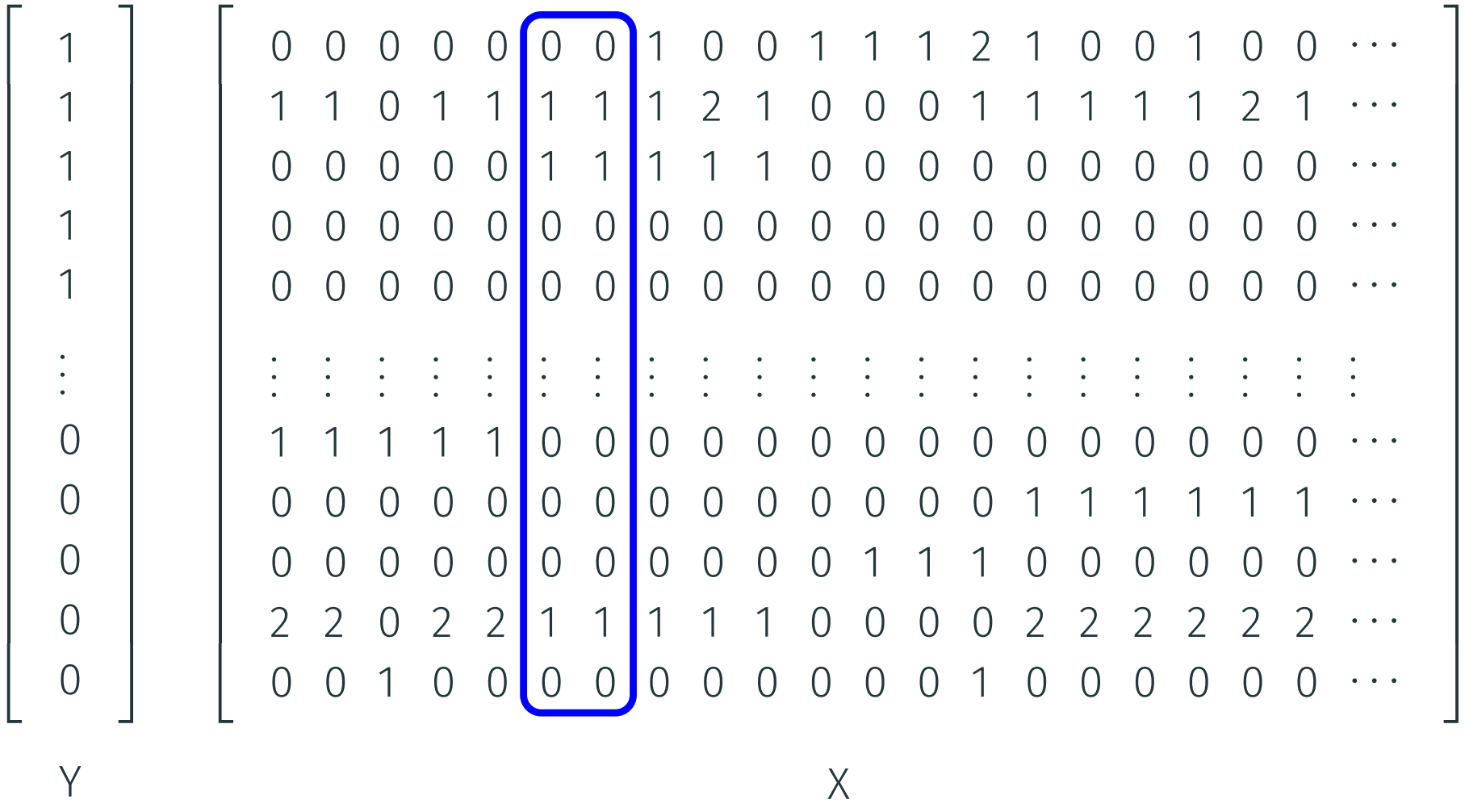
Genomewide association studies: a brief introduction

$$
\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\quad
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
2 & 2 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & \cdots \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots
\end{bmatrix}
$$

Y

X

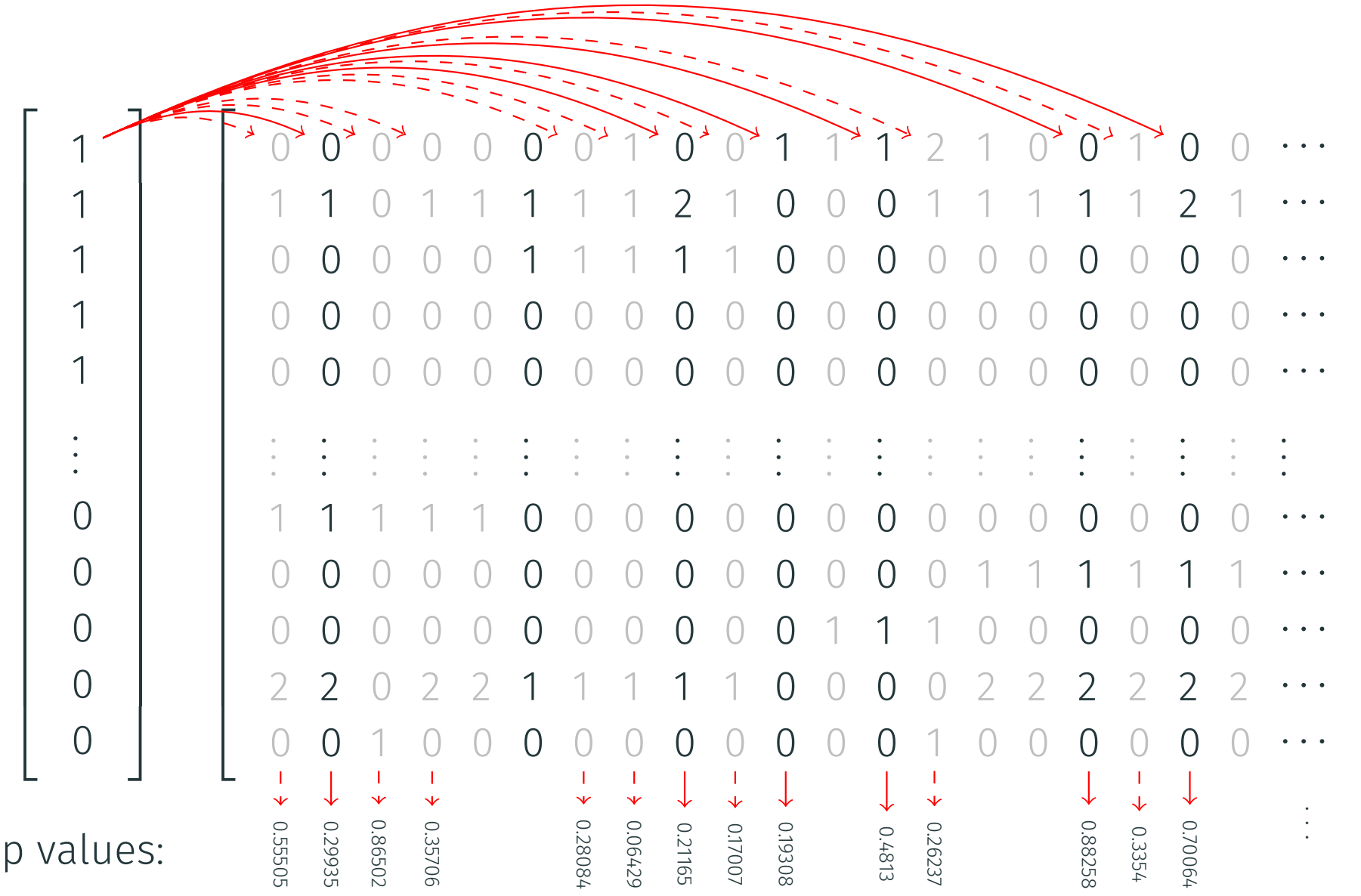
Genomewide association studies: a brief introduction



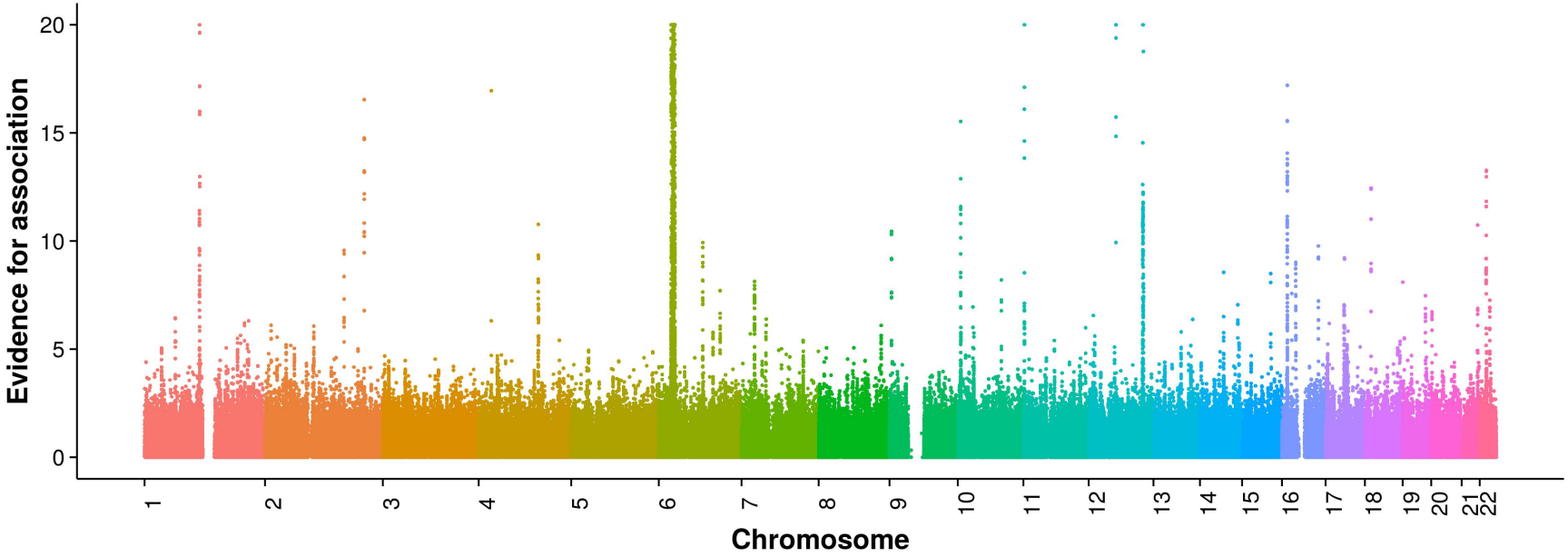
Genomewide association studies: a brief introduction

Genomewide association studies: a brief introduction

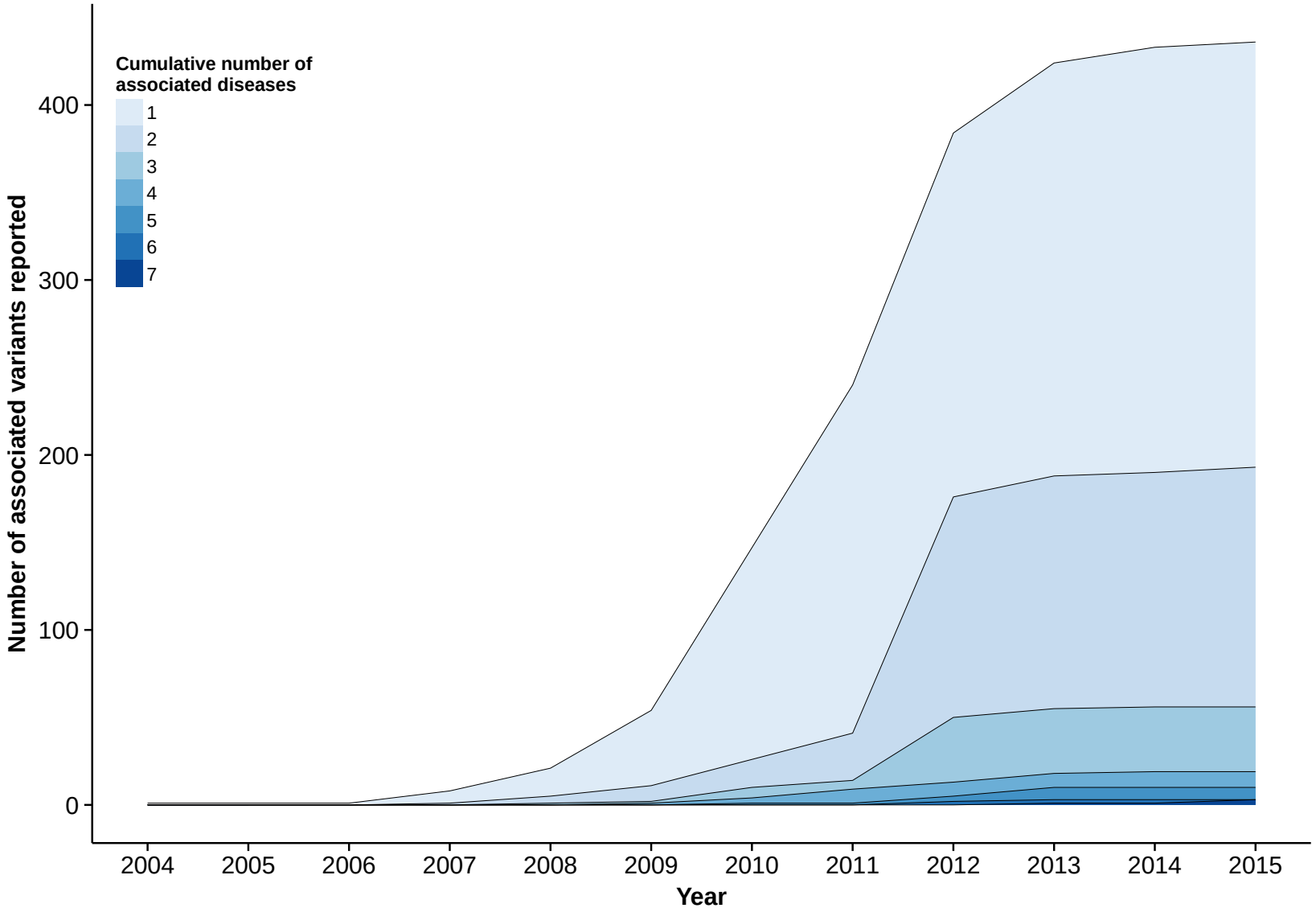
Genomewide association studies: a brief introduction



Manhattan plots give a visual summary of results

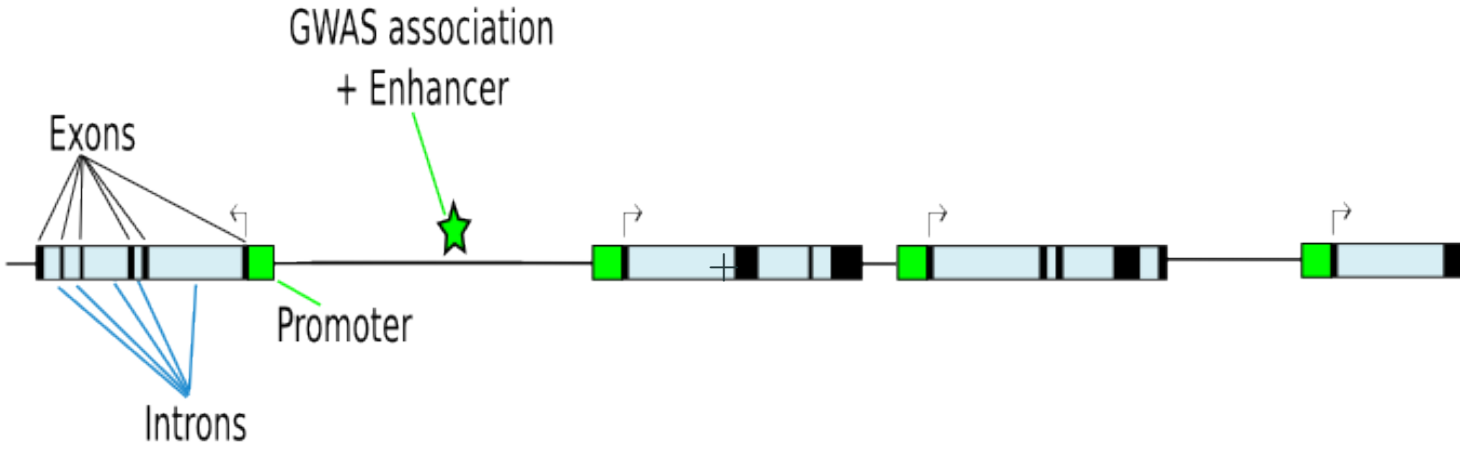


Reports of genetic associations in 19 autoimmune diseases

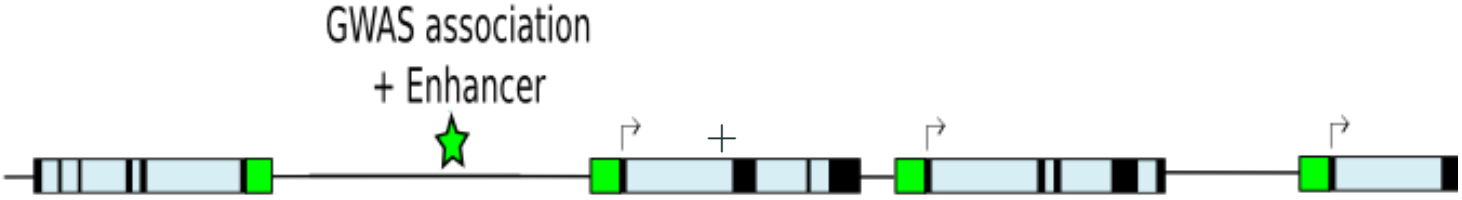


Inferring disease relevant cells

GWAS disease variants tend to be in cell specific enhancers

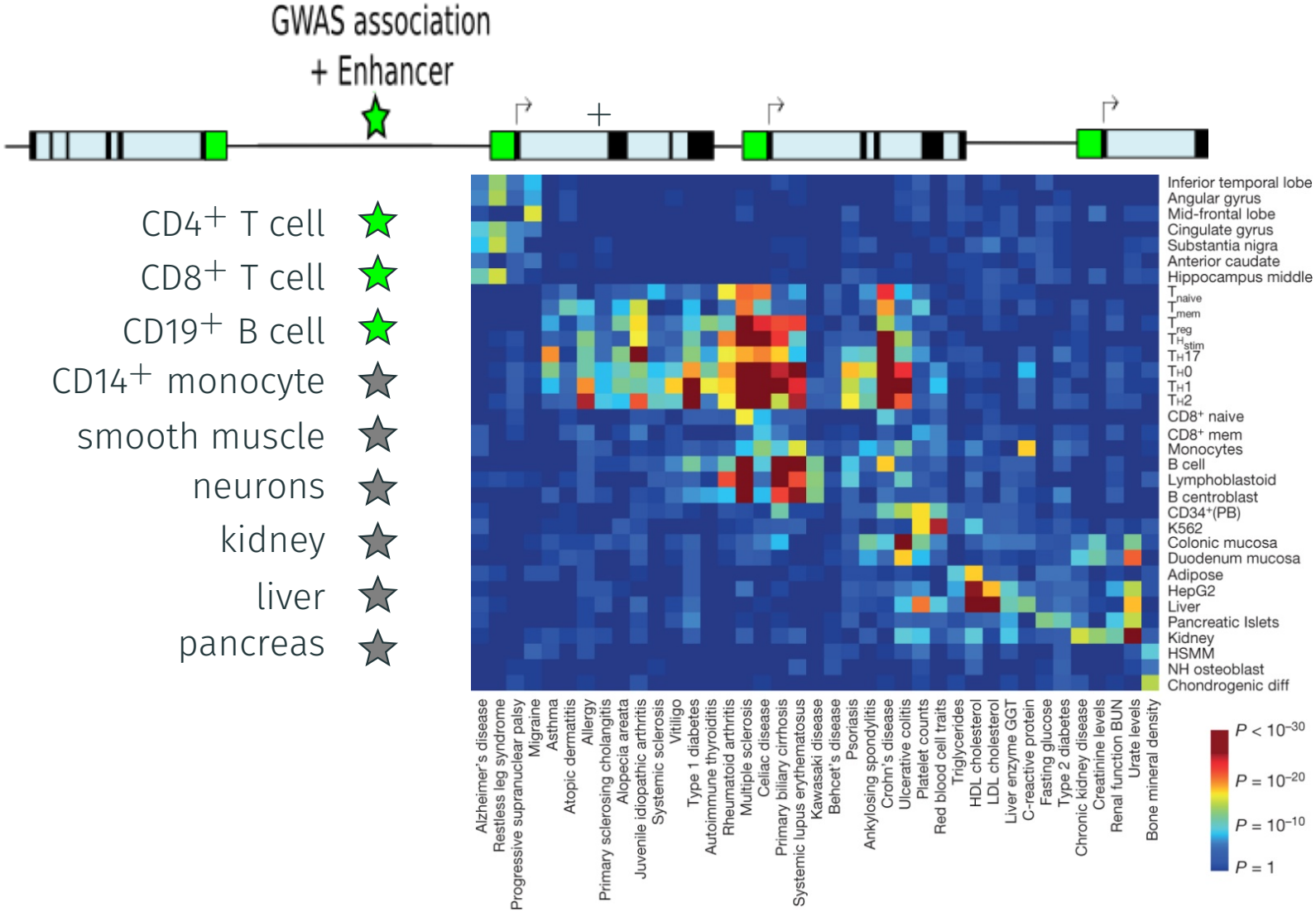


GWAS disease variants tend to be in cell specific enhancers



CD4$^+$ T cell ★
CD8$^+$ T cell ★
CD19$^+$ B cell ★
CD14$^+$ monocyte ★
smooth muscle ★
neurons ★
kidney ★
liver ★
pancreas ★

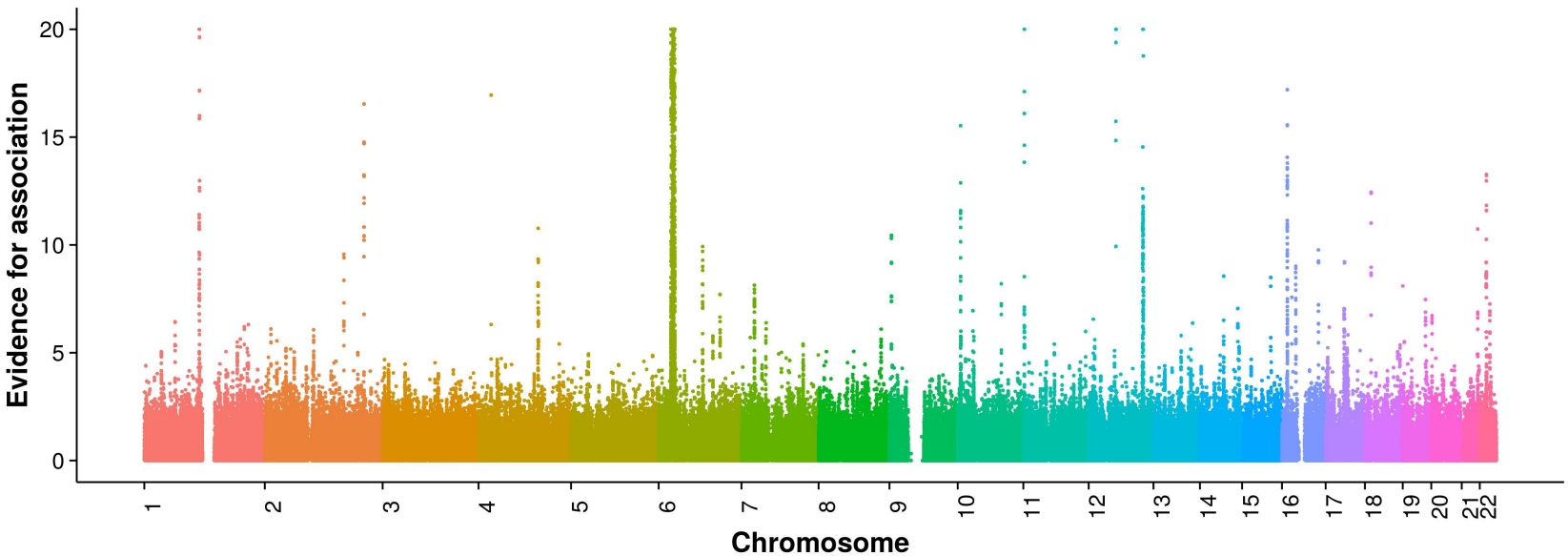
GWAS disease variants tend to be in cell specific enhancers



Farh et al., Nature 2014

Mapping causal variants

Association does not identify the causal variant



Fine mapping causal variants

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”

Stepwise regression

1. Find the single SNP $j^{(1)}$ which minimises the objective function

$$\sum_i (y_i - \beta_j x_{ij})^2$$

2. Find the next SNP $j^{(2)}$ which minimises the objective function

$$\sum_i (y_i - \beta_1 x_{ij^{(1)}} - \beta_j x_{ij})^2$$

3. Continue until the change in the objective function is "non-significant"

Stepwise methods find **one** solution. Every addition is conditional on what has been added at previous steps.

Simplified example of stepwise search

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

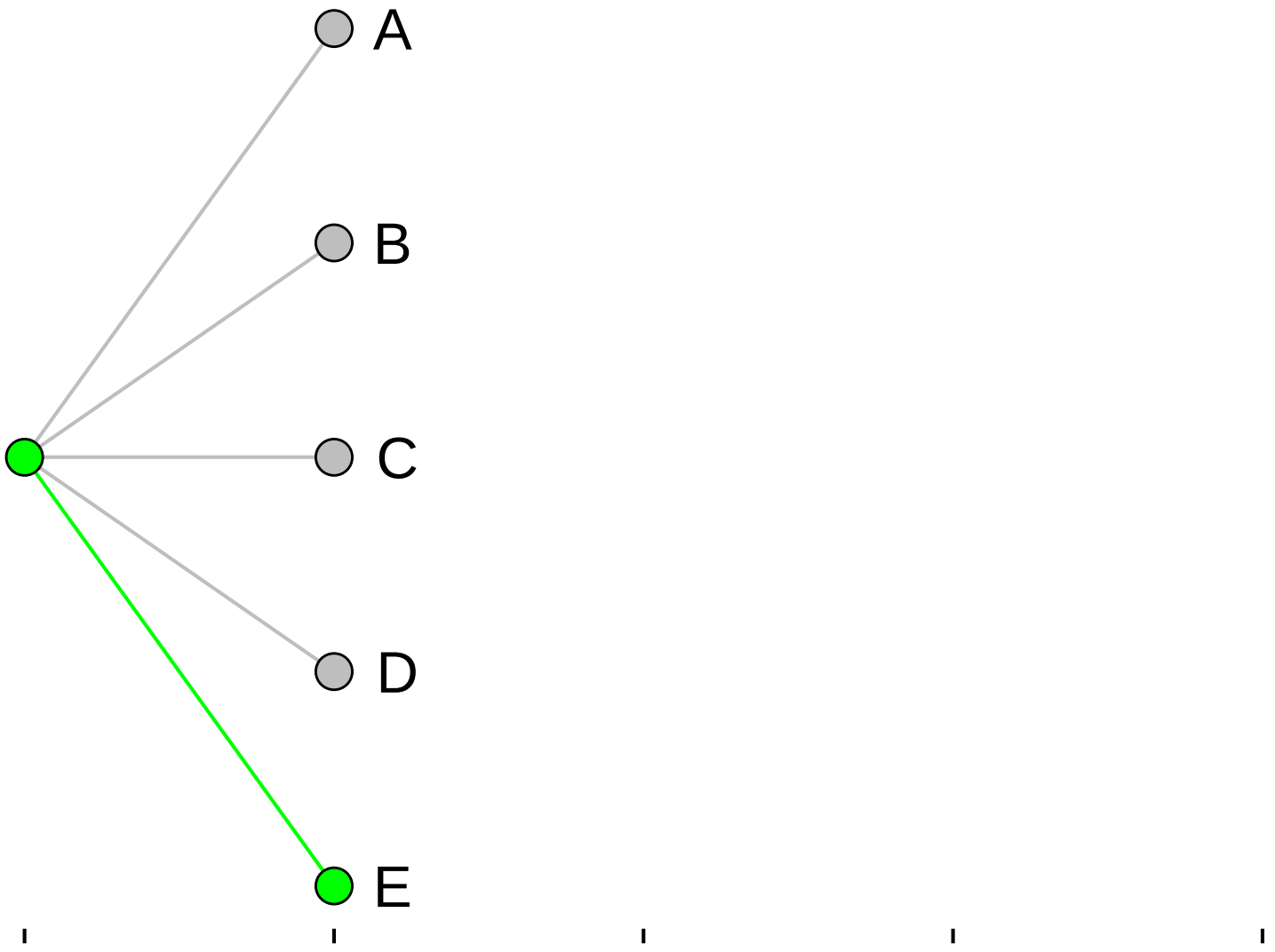
Consider ”models”, combinations of SNPs:

- A
- B + C

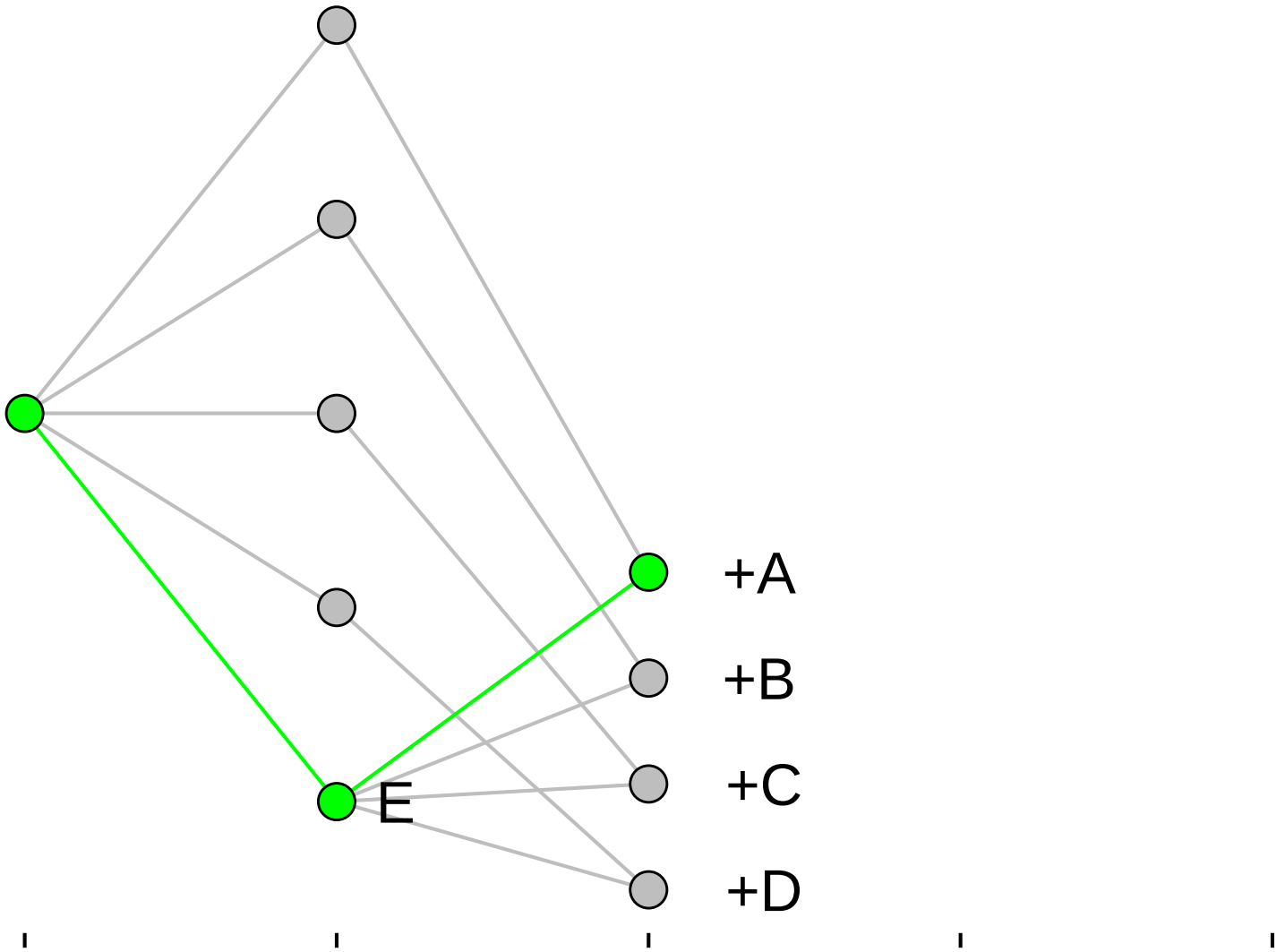
There are 32 possible models. How can we find the ”best” one?

1: start with no SNPs

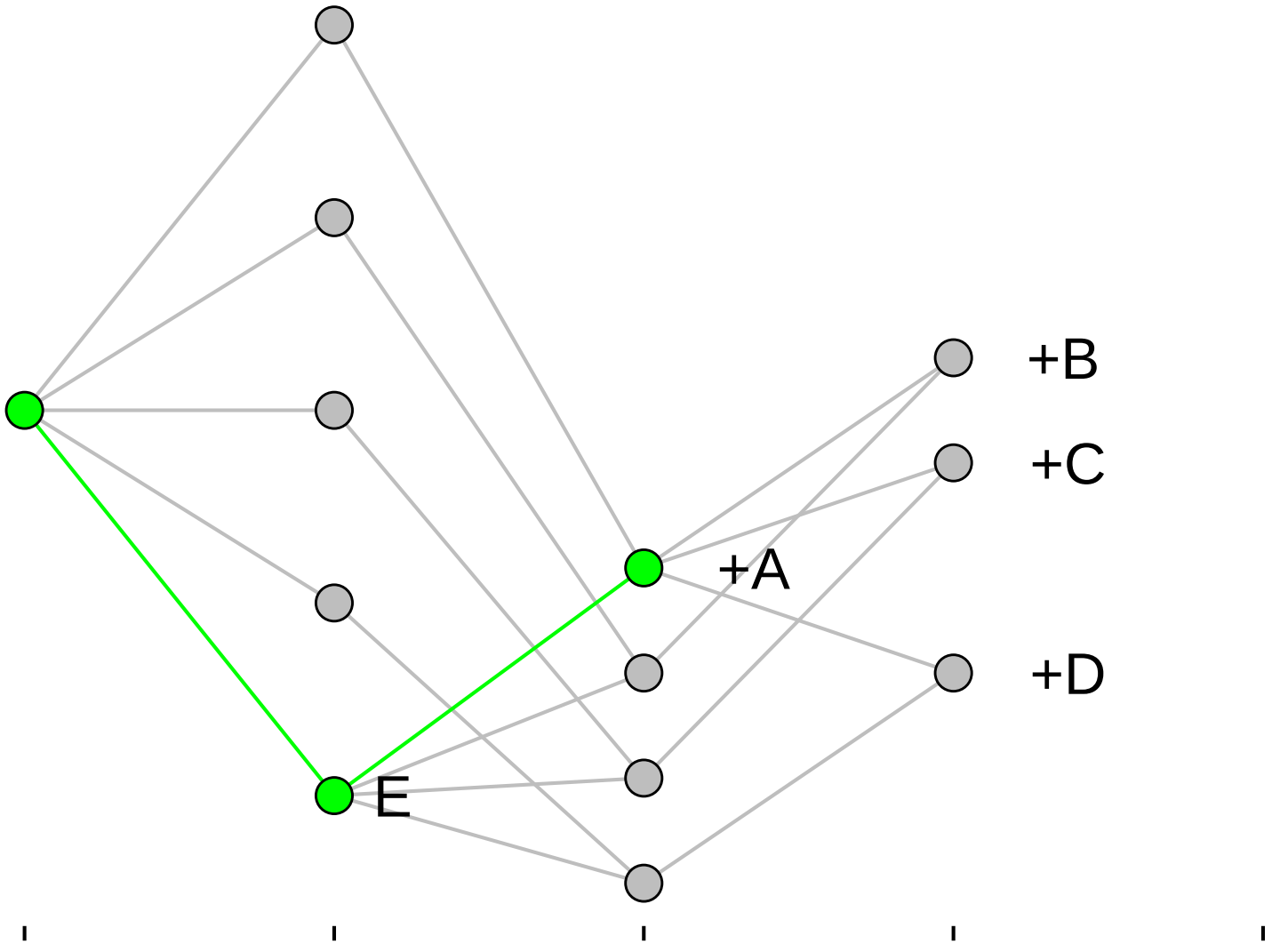
2: best model assuming single causal variant



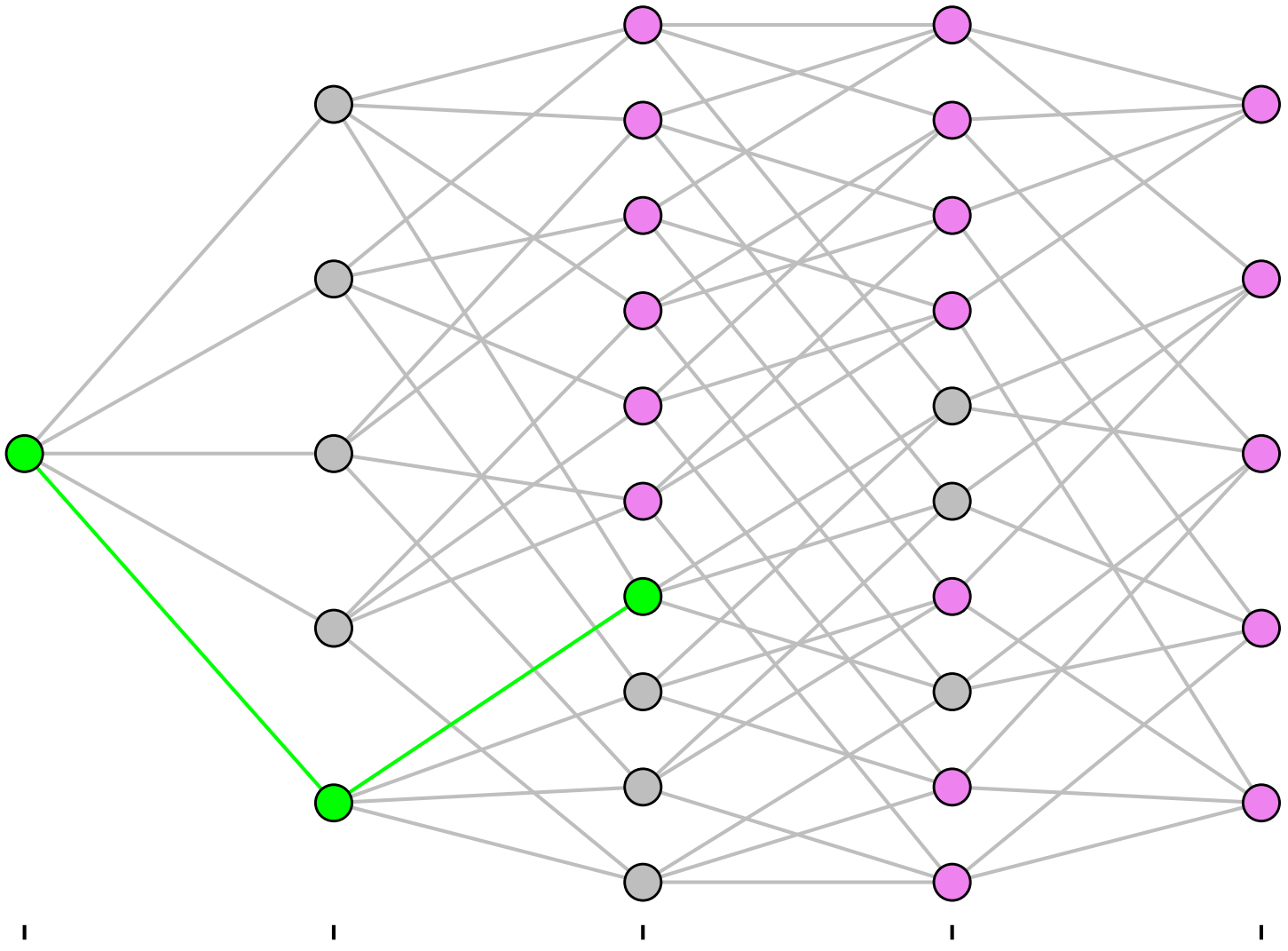
3: step forward into two SNP models



4: consider three SNP models



We only explored a subset of models



Alternative: regularized regression

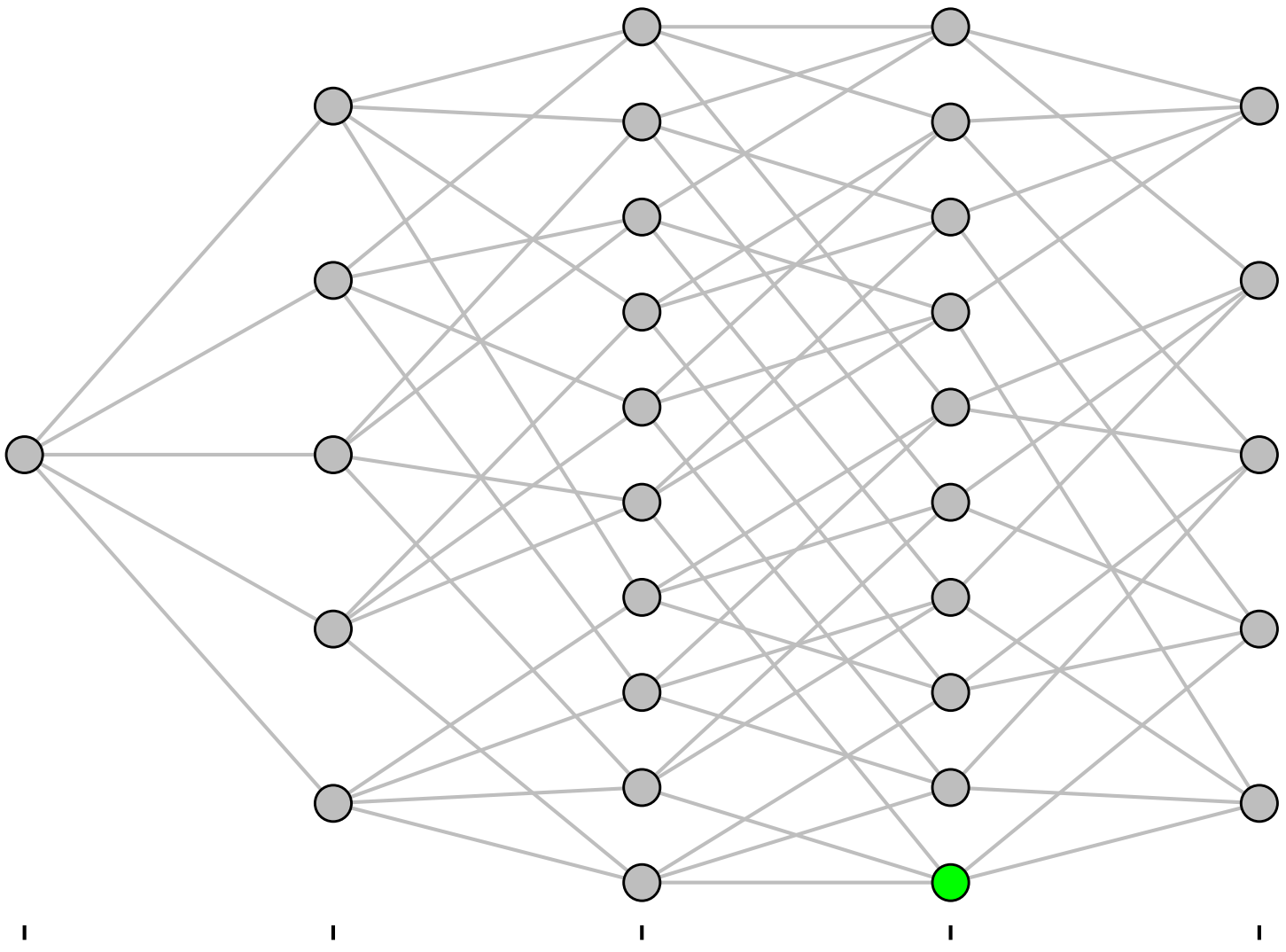
Lasso regression

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

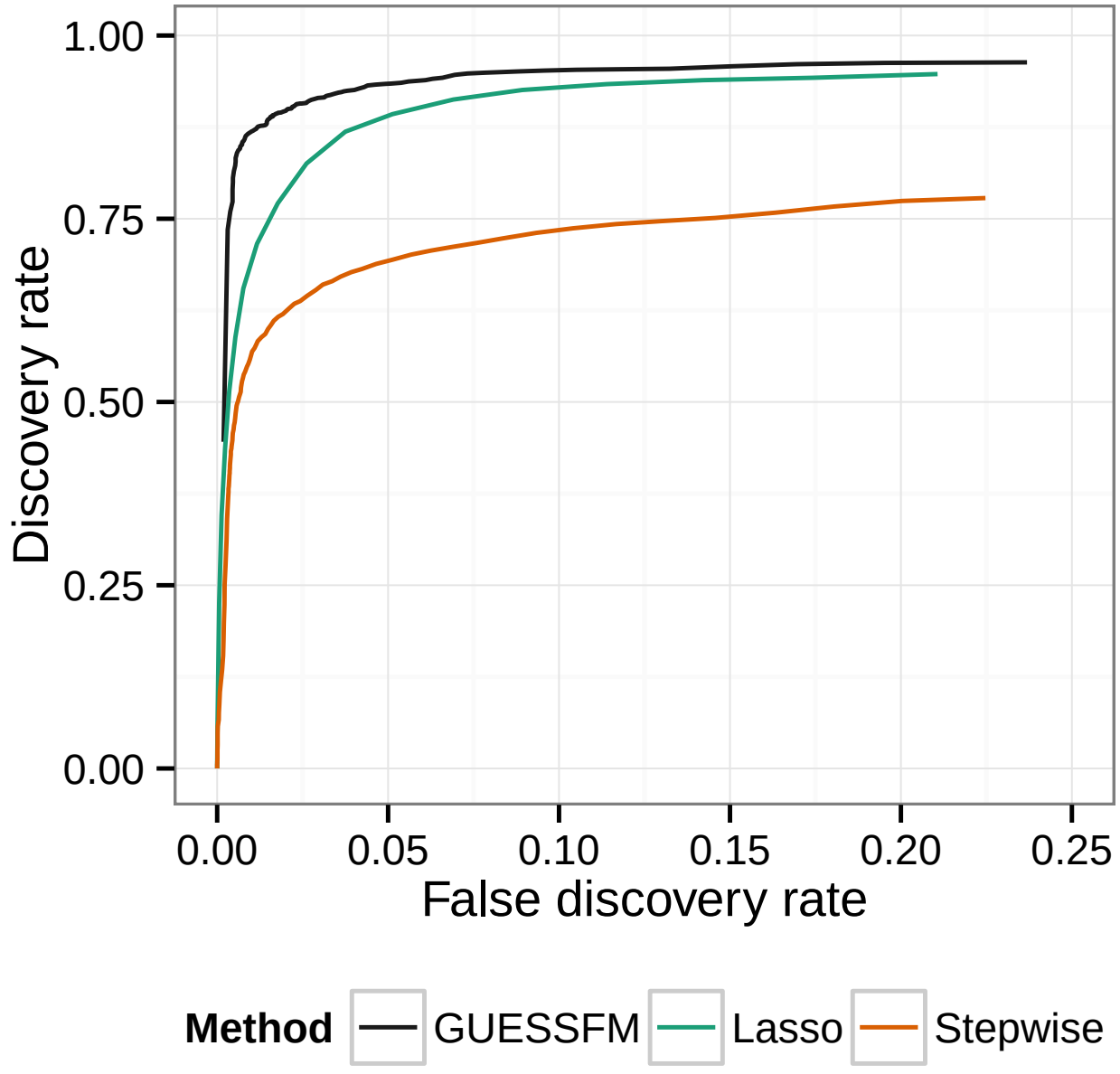
$$\sum_i (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

λ is typically chosen by cross validation.

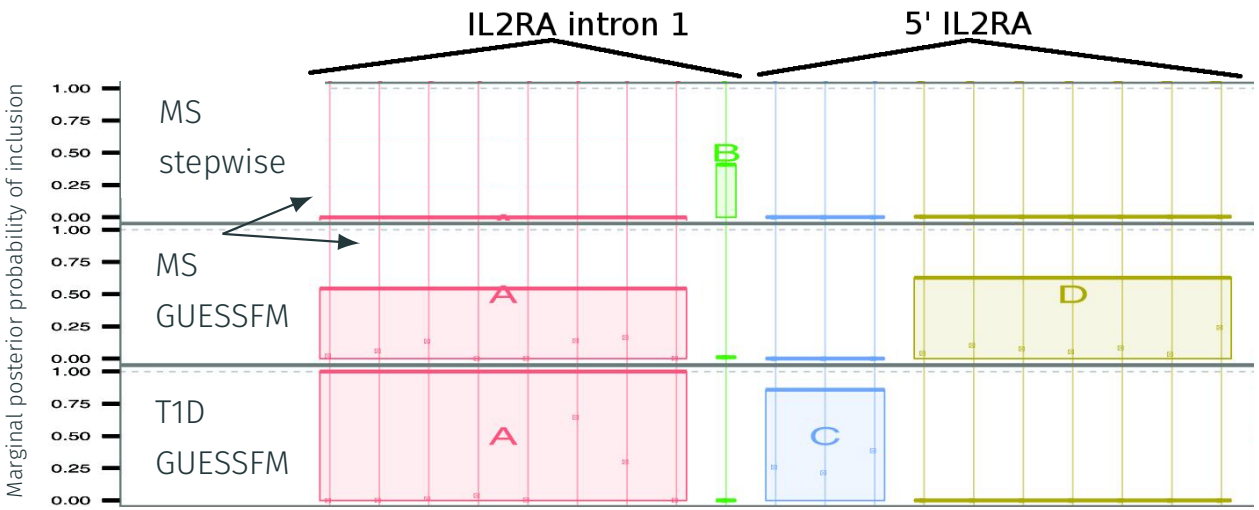
Alternative (2): ‘stochastic search’ visits all important models



Better recovery of “causal” variants in simulated data



Better recovery of “causal” variants in real data

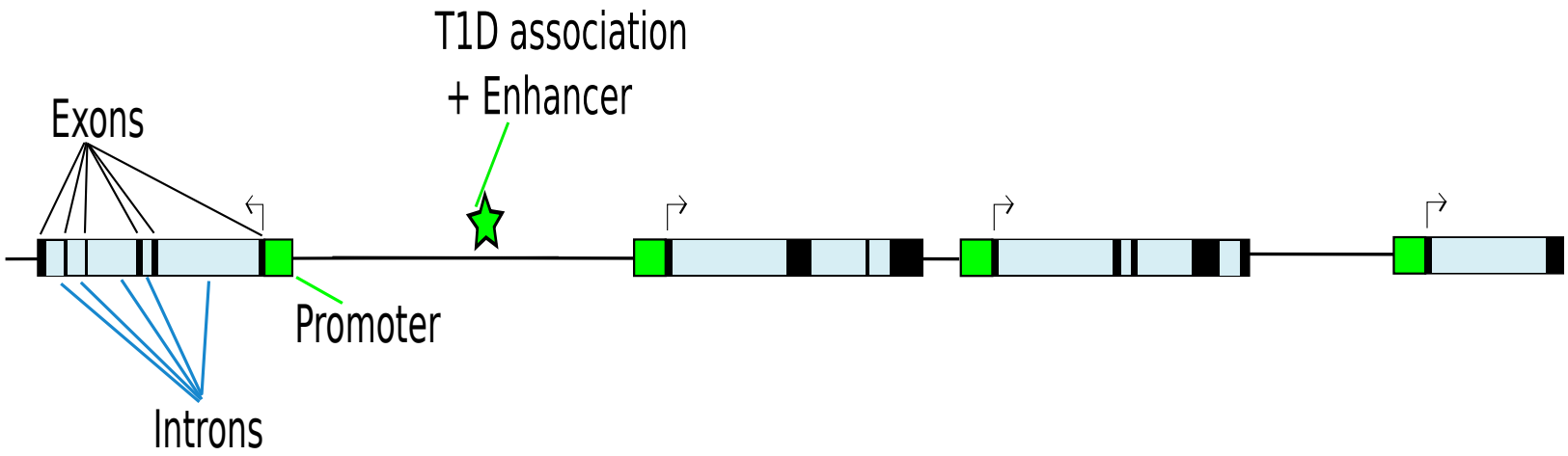


Haplotype			Fq (%)	OR	95%–CI	p
A	D	B				
A	G	T	69.13	1.00	–	–
A	**A**	**C**	15.05	0.80	0.76–0.84	$< 2 \times 10^{-16}$
G	G	**C**	9.99	0.82	0.78–0.86	4.66×10^{-13}
A	**A**	T	5.59	0.85	0.79–0.91	8.06×10^{-06}

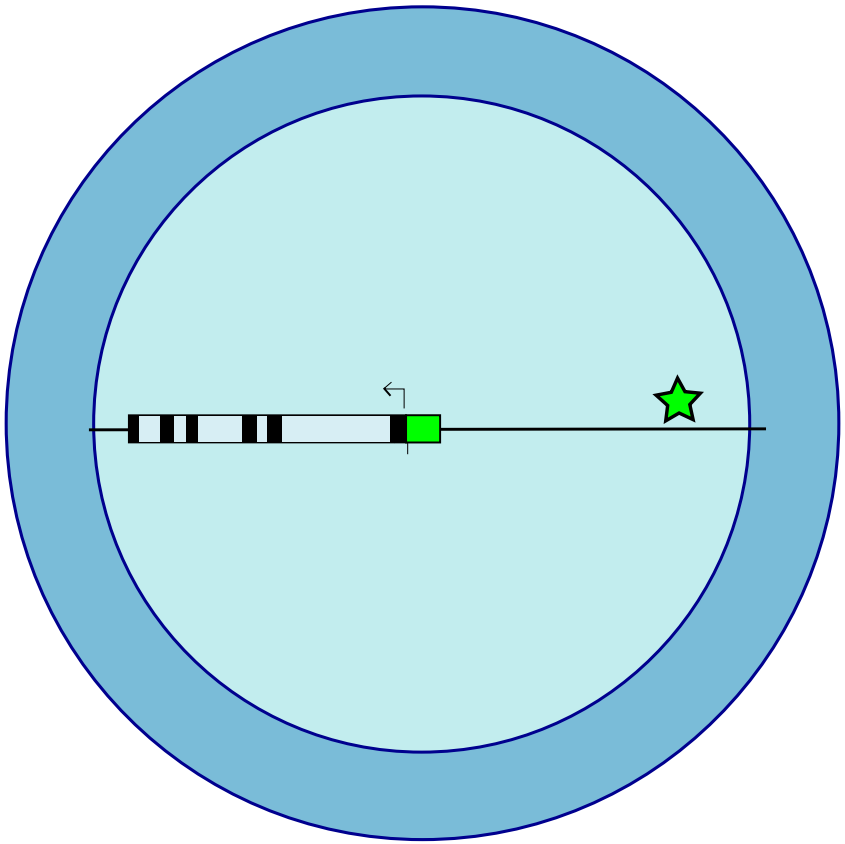
Wallace et al, PLOS Genet 2015

Identifying causal genes from causal variants

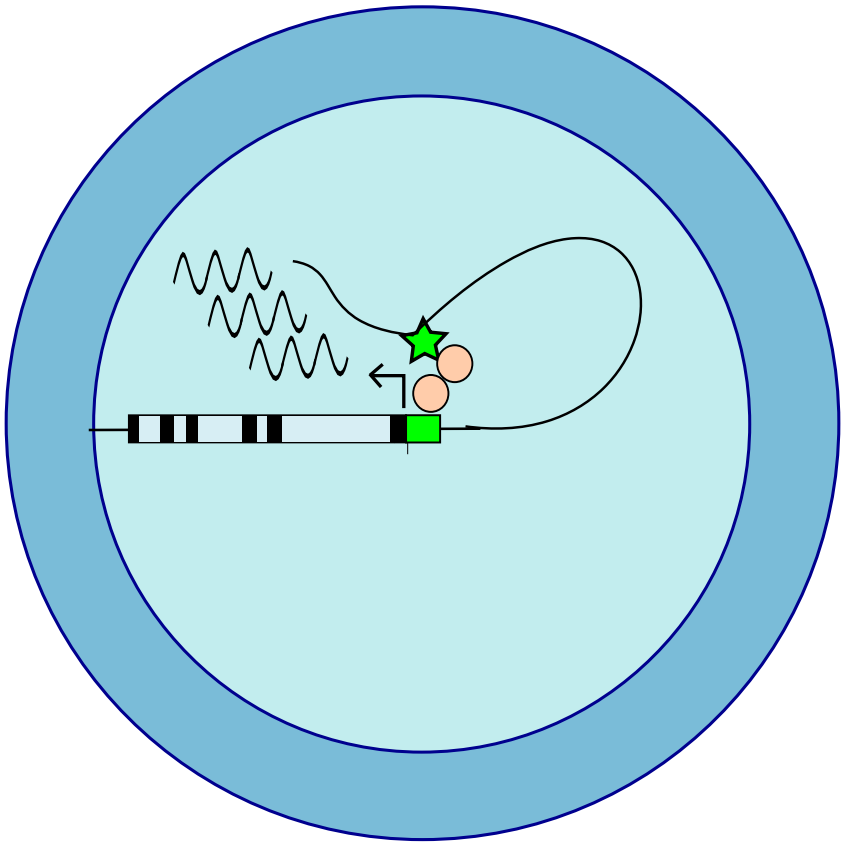
Disease causal variants do not directly implicate causal gene



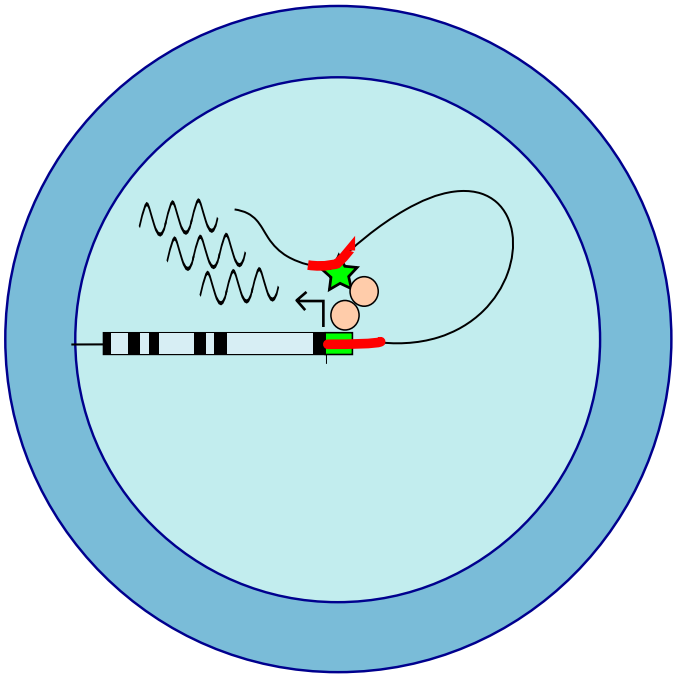
Disease causal variants do not directly implicate causal gene



3D folding of DNA in nucleus connects enhancers to promoters



Promoter capture Hi-C in 17 primary cell types



Endothelial precursors

Erythroblasts

Neutrophils + precursors

Megakaryocytes

Macrophages (M0,M1,M2)

CD4+ T cells, naïve and total

CD8+ T cells, naïve and total

Monocytes

Fetal thymus

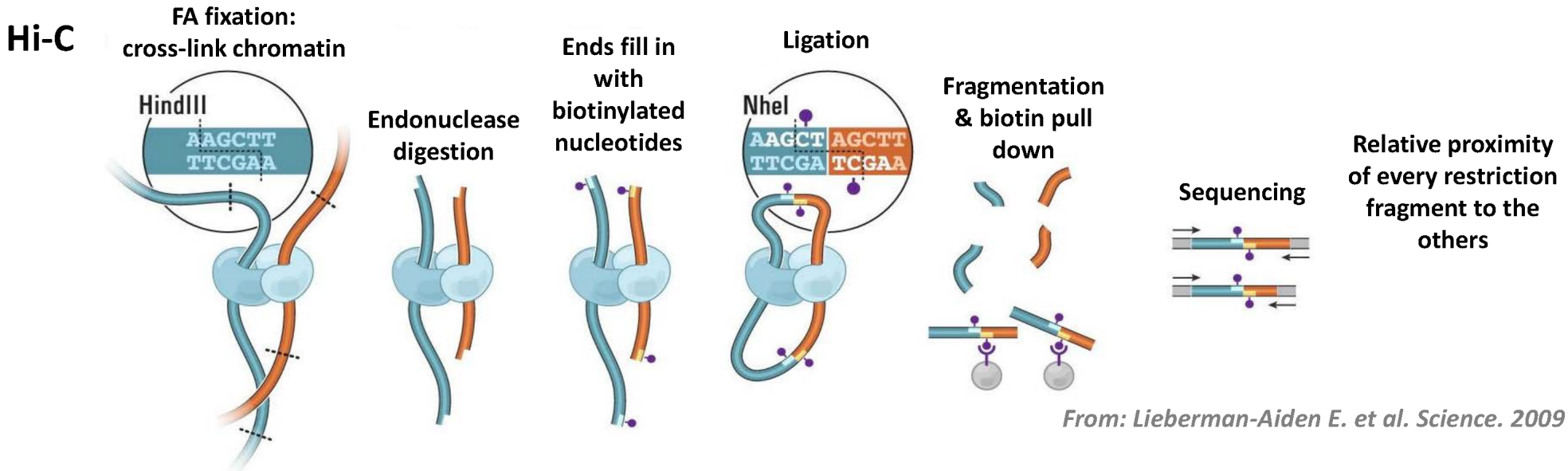
B cells, naïve and total

Total CD4+, activated, non-activated

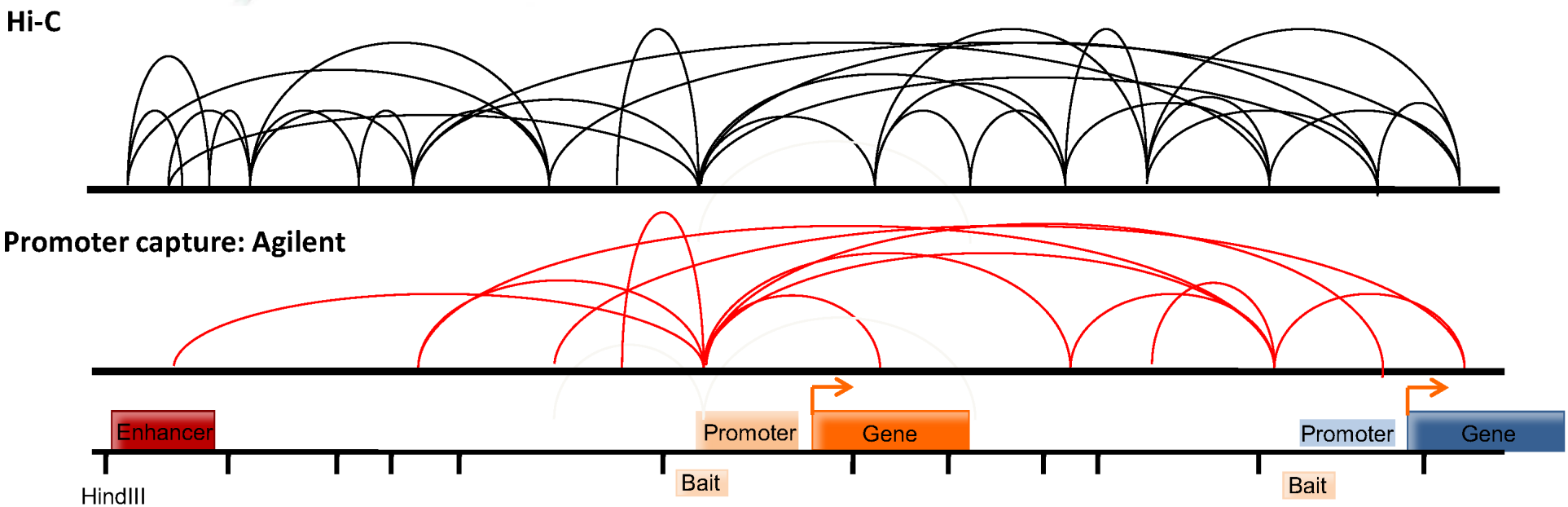

Hi-C
FLASHIN'
FRUIT PUNCH

Analysing genome wide architecture in 3-dimensions : Hi-C

Assigning regulatory regions to their putative target genes



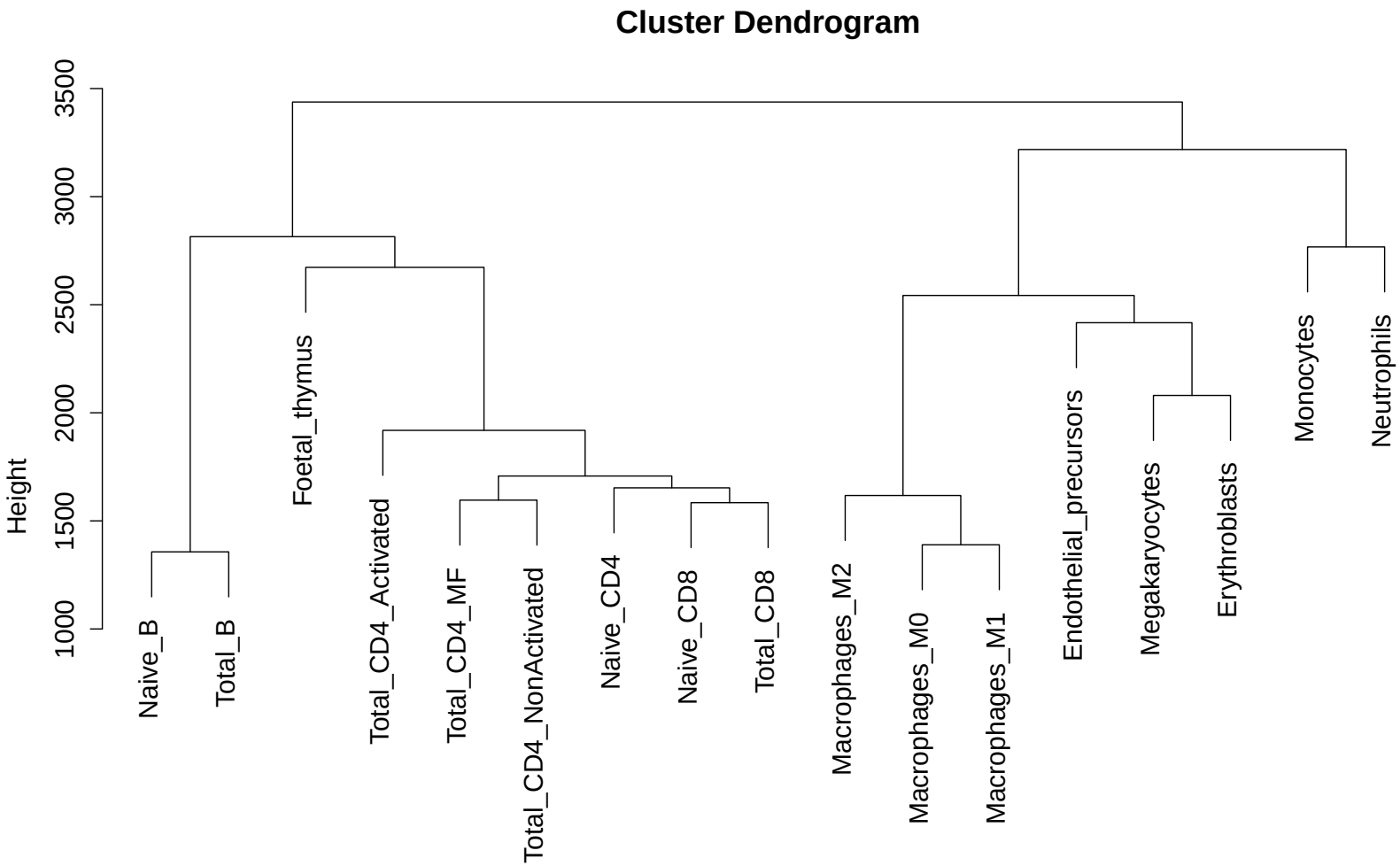
From: Lieberman-Aiden E. et al. Science. 2009



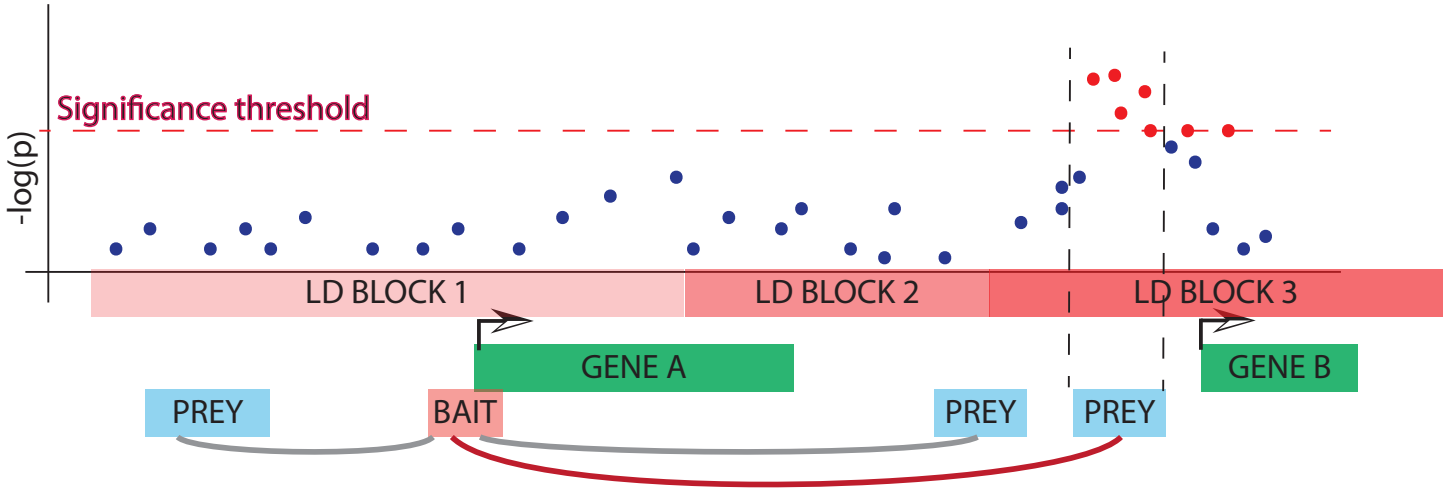
39,460 probes covering 22,459 Hind III promoter fragment
(18,868 mRNA promoters, 1026 microRNA genes, 1171 snRNA, 1099 snoRNA, 232 Ultra Conserved Elements)

Biola Javierre

PCHi-C maps are lineage specific



Using PCHi-C to inform interpretation of GWAS



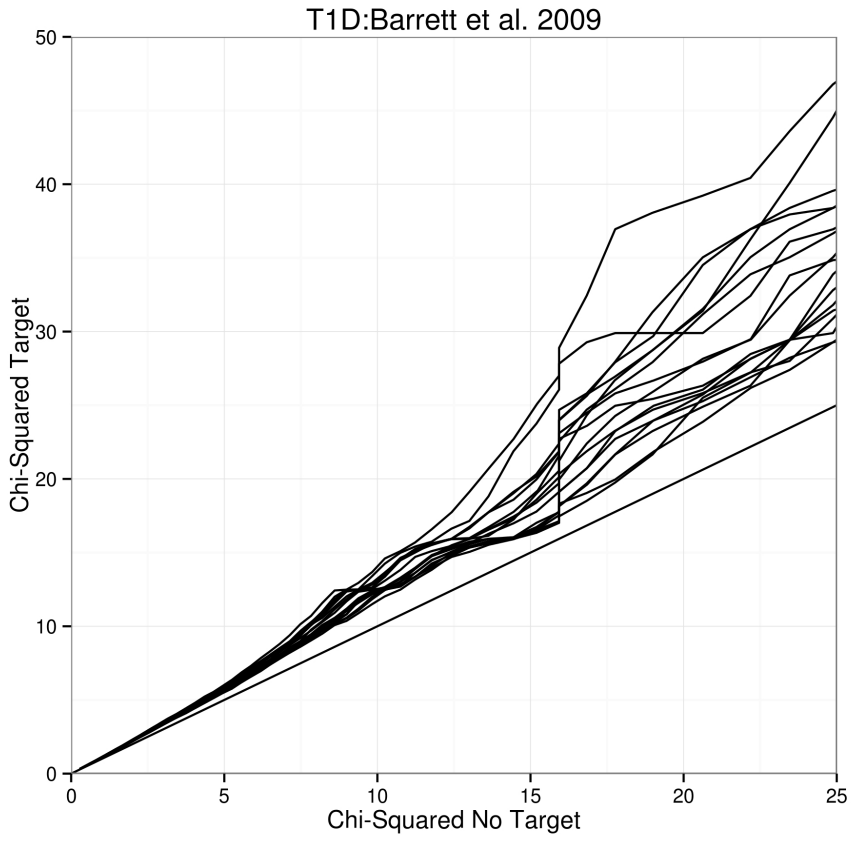
Enrichment analysis

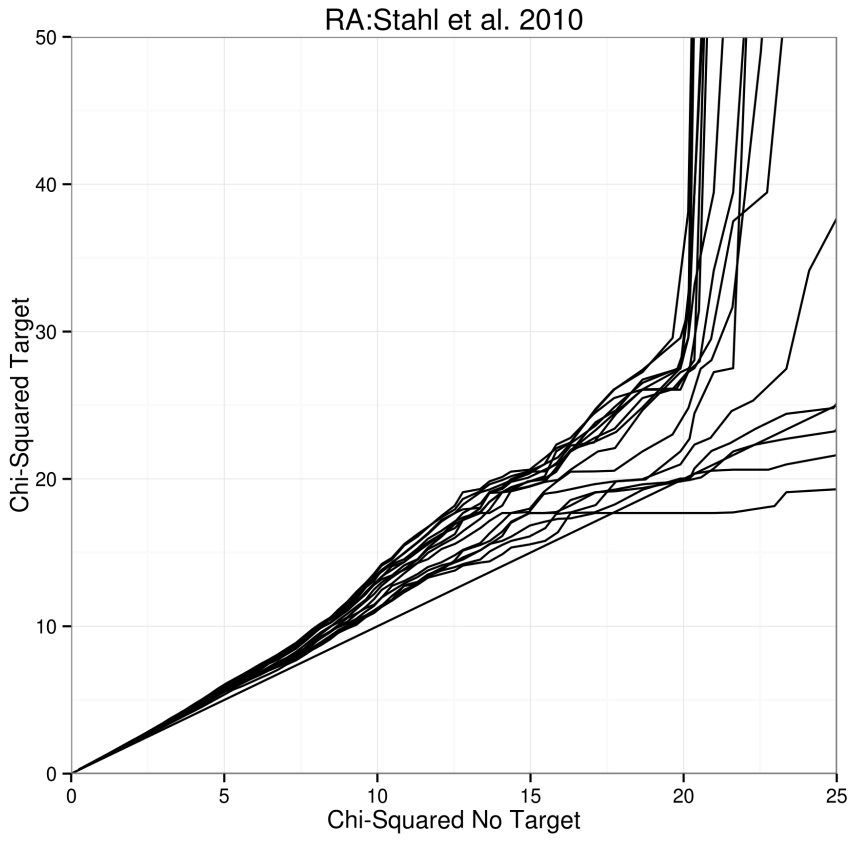
- Do GWAS signals differ between baits from different cell types?

Causal gene lookup

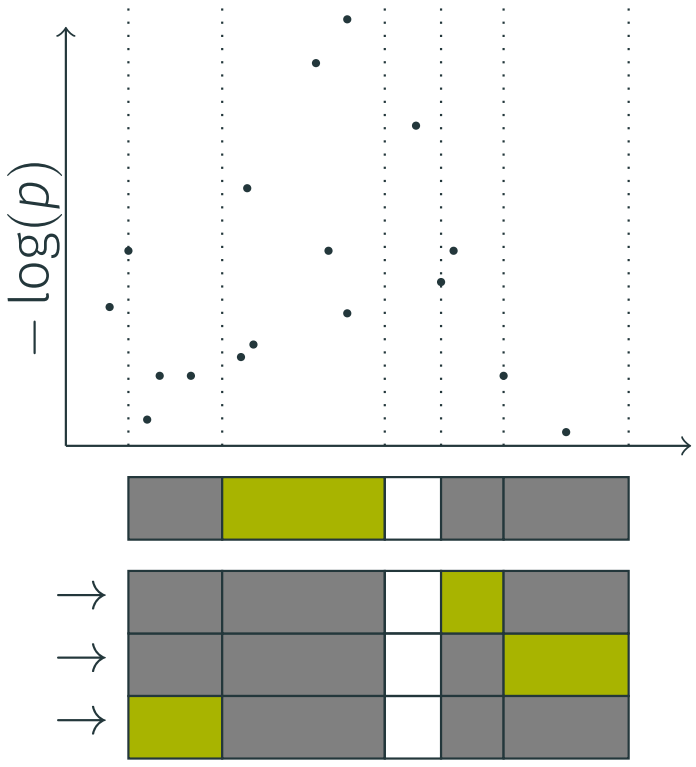
- If we fine map disease causal variant(s), can we use PCHi-C to 'lookup' the causal gene?

GWAS enrichment varies by cell type





Testing relative enrichment of GWAS signals

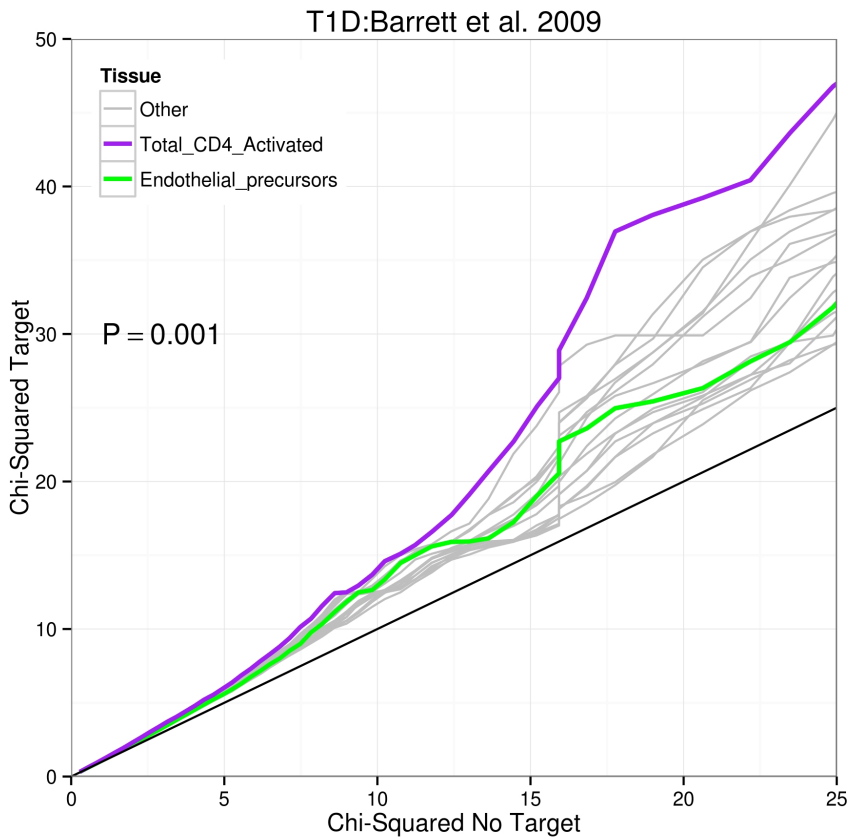


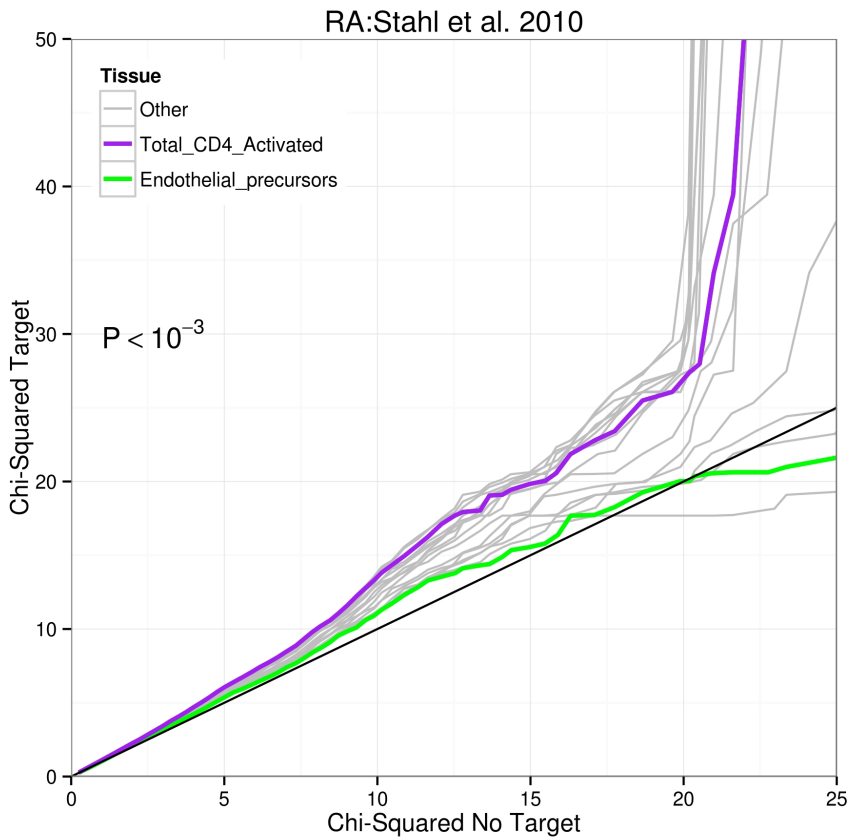
1. Generate superblocks of contiguous HindIII fragments with CHIC annotation *allow single fragment gaps*
2. Compare p values in green regions *versus* grey, eg difference in mean$(-\log(p))$
3. Generate null distribution by also calculating mean difference in all shifted patterns

Aim to preserve local correlation structure between SNPs (p values) *and* between regulatory sequences.

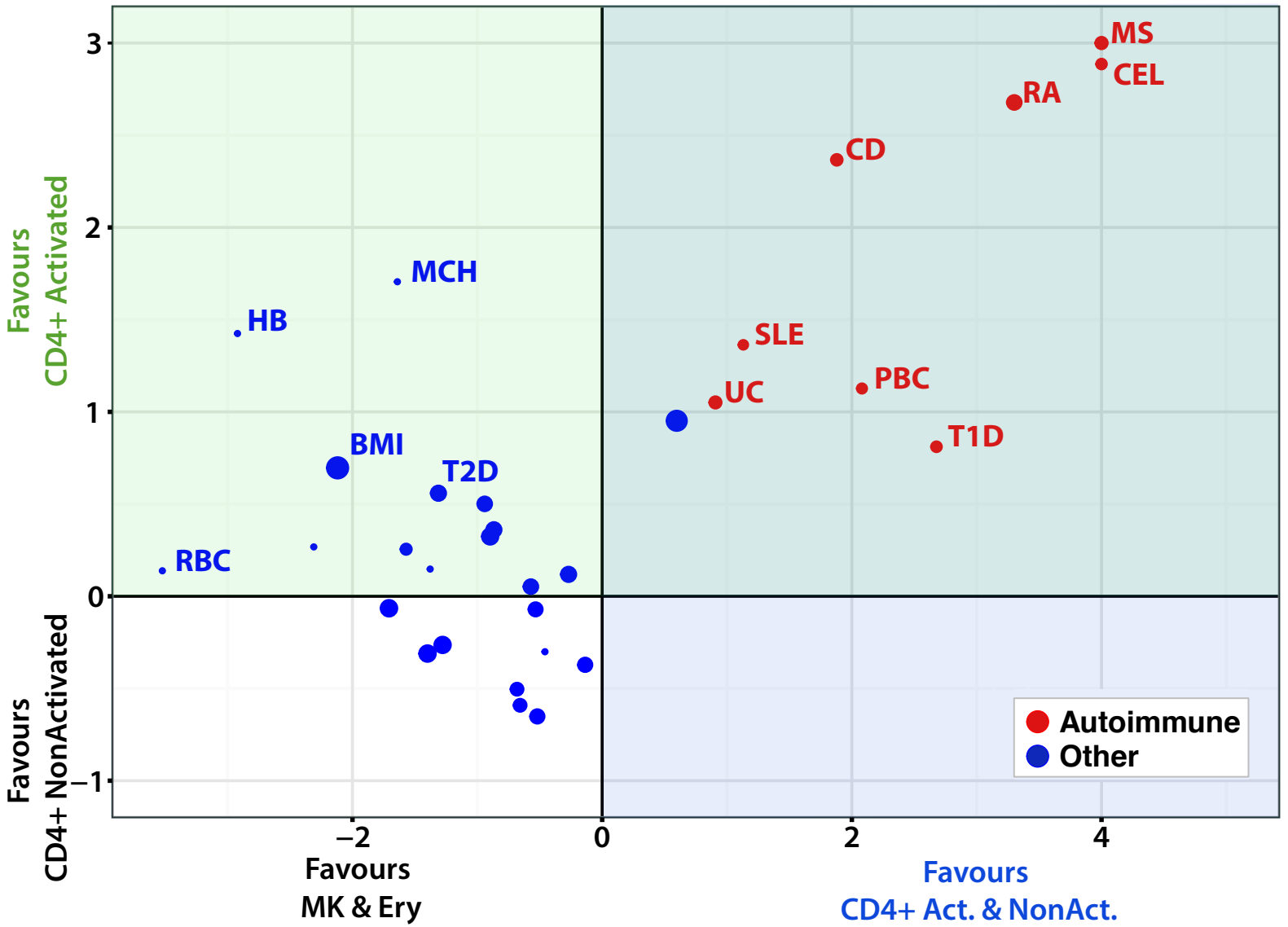
Limited number of combinations at any "superblock", but many superblocks $\rightarrow$ fast simulation of null distribution

Enrichment significantly greater in activated CD4 vs endothelial precursors

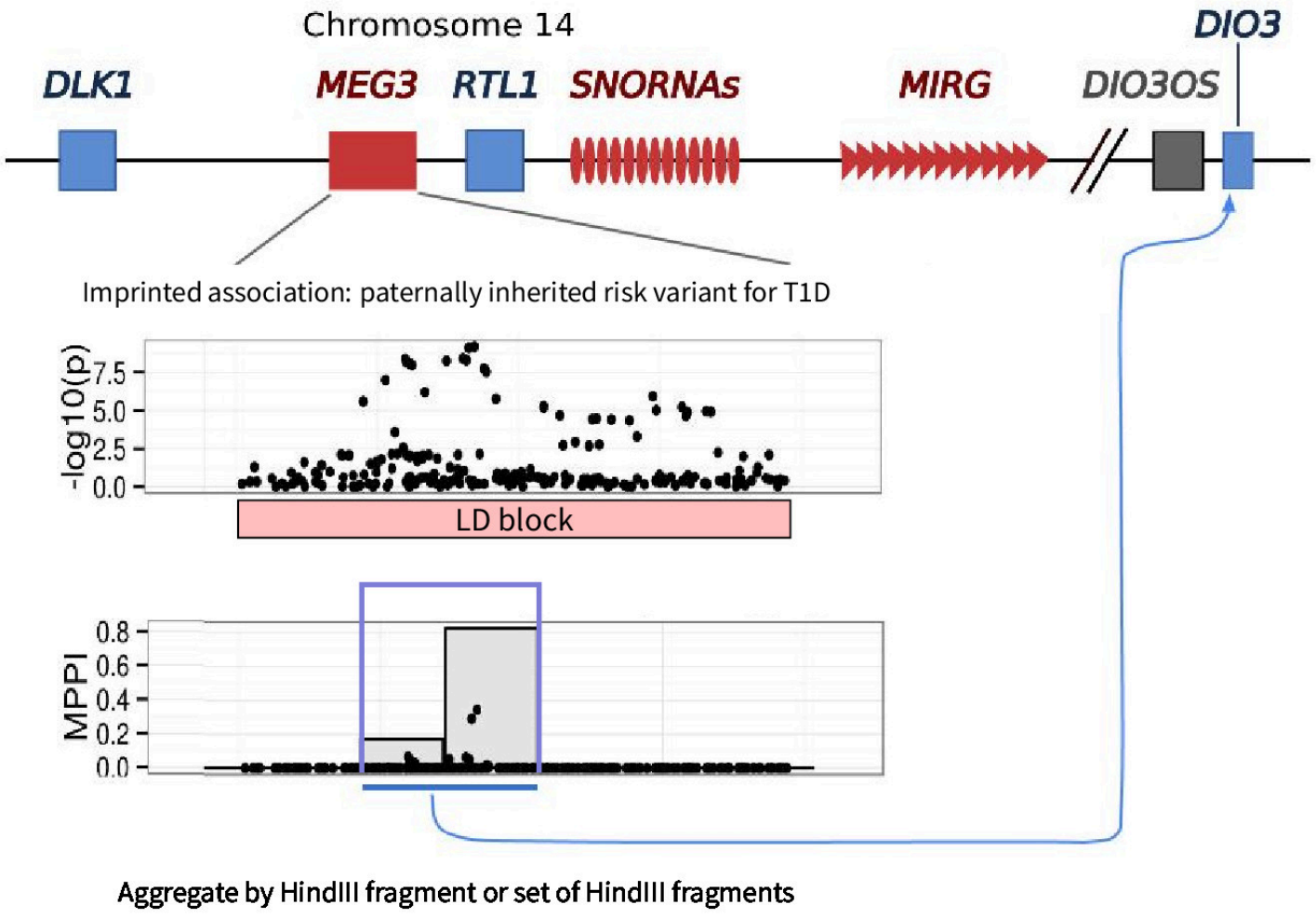




GWAS signals are enriched in PCHi-C interacting regions



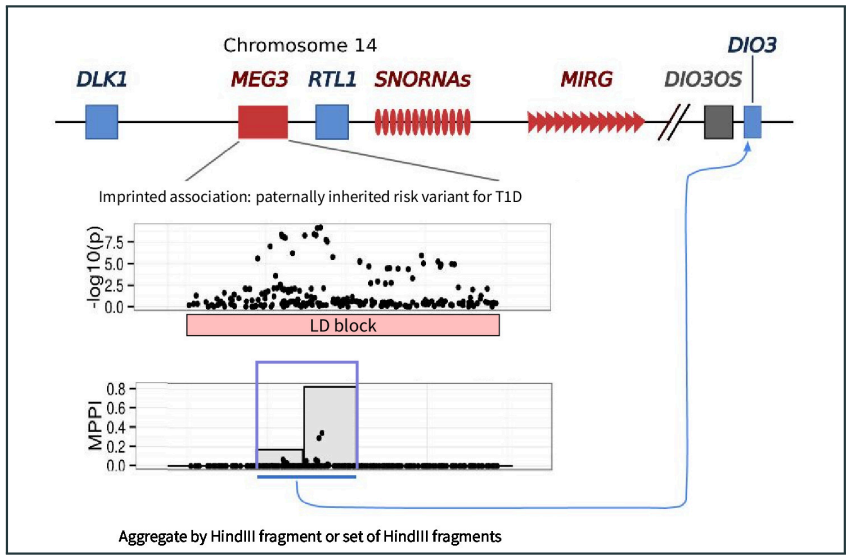
Fine mapping + CHi-C analysis of T1D region on chromosome 14q32



Nat Genet 2010

The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes

Chris Wallace, Deborah J Smyth, Meeta Maisuria-Armer, Neil M Walker, John A Todd & David G Clayton



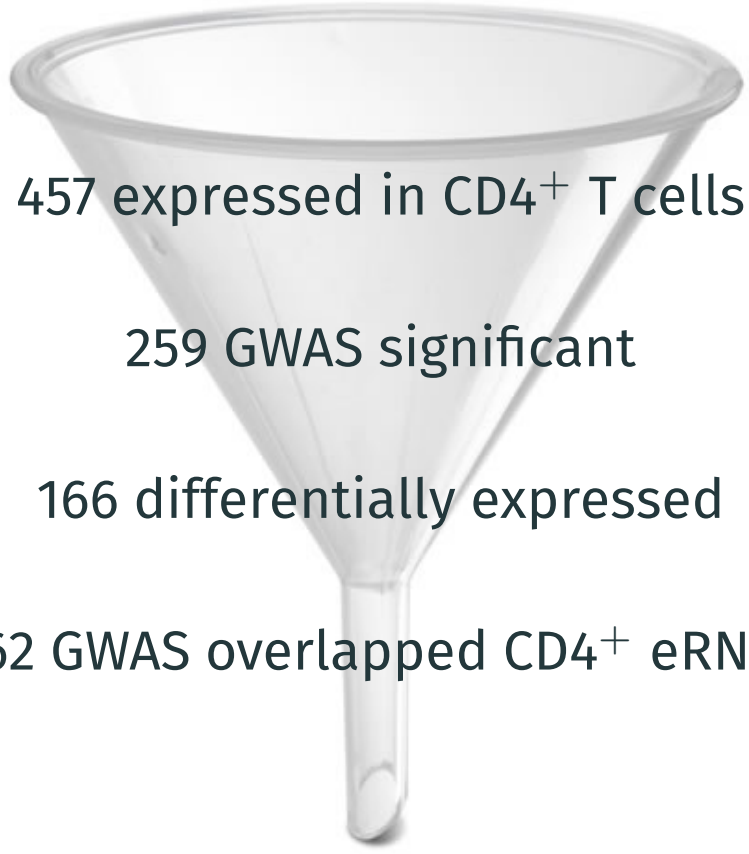
The Thyroid Hormone-Inactivating Type III Deiodinase Is Expressed in Mouse and Human β-Cells and Its Targeted Inactivation Impairs Insulin Secretion

Mayrin C. Medina, Judith Molina, Yelena Gadea, Alberto Fachado, Monika Murillo, Gordana Simovic, Antonello Pileggi, Arturo Hernández, Helena Edlund, and Antonio C. Bianco

Endocrinol 2011

Framework generates a long list of results

602 autoimmune disease prioritised genes in $CD4^+$ T cells



457 expressed in $CD4^+$ T cells

259 GWAS significant

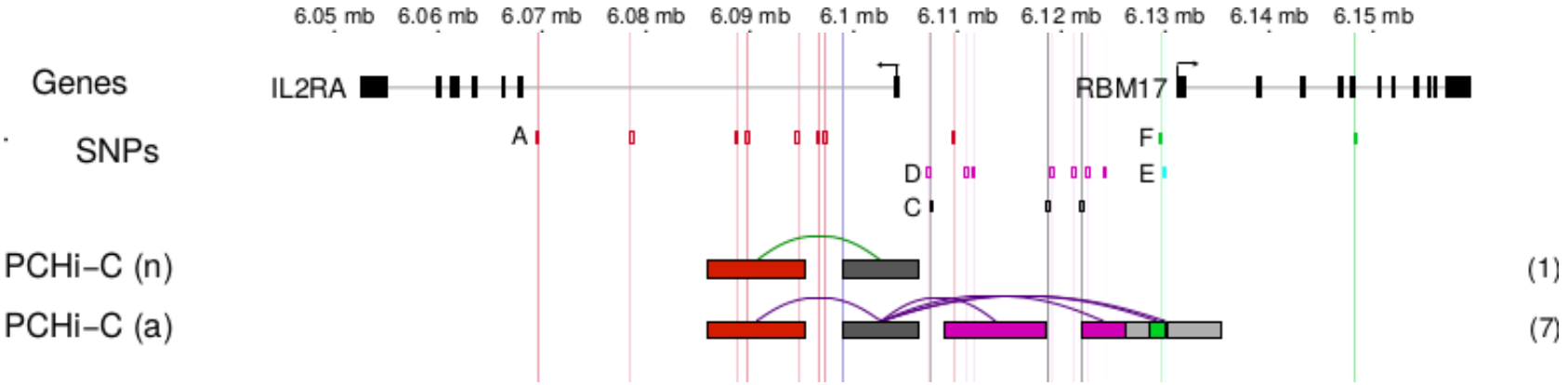
166 differentially expressed

62 GWAS overlapped $CD4^+$ eRNA

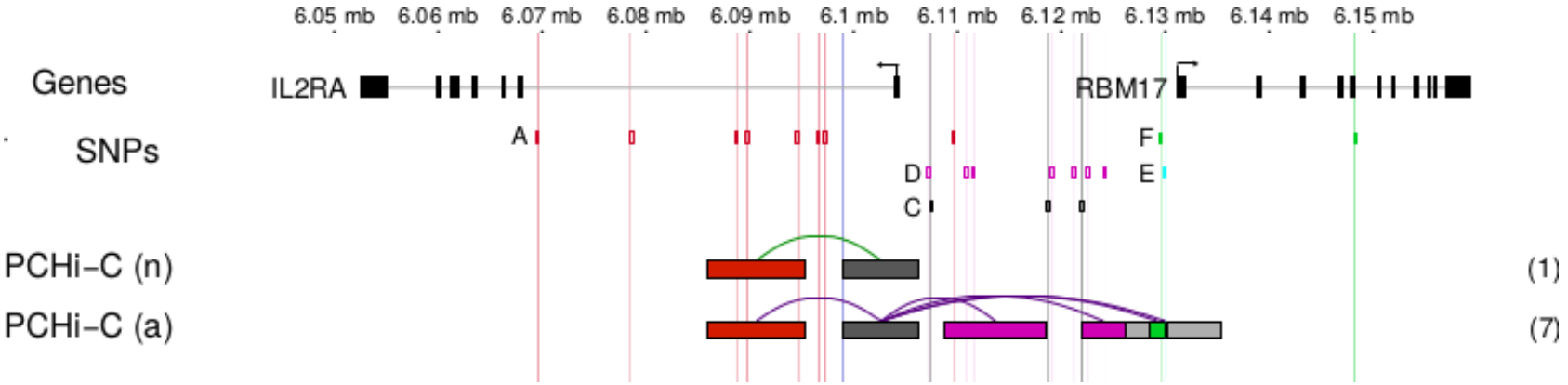
Require functional validation experiments

IL2RA prioritised in activated and non-activated CD4^{+} T cells

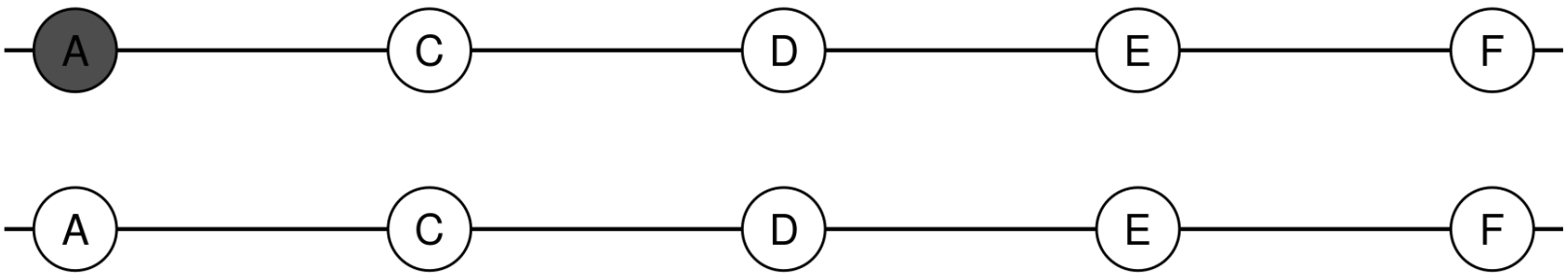
IL2RA prioritised in activated and non-activated CD4^{+} T cells



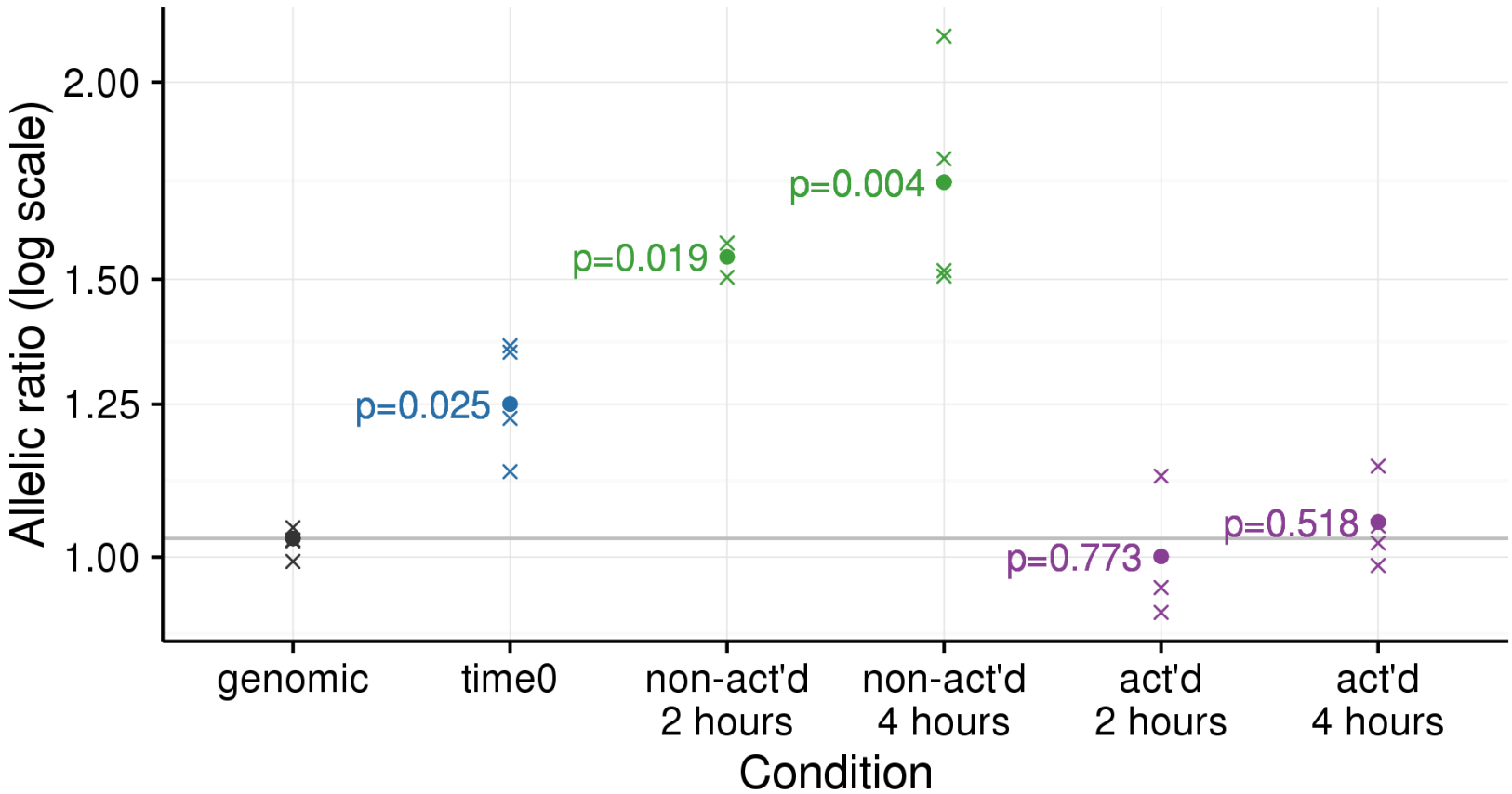
IL2RA prioritised in activated and non-activated CD4$^+$ T cells



Allele specific expression: quantify relative expression of two chromosomes using targetted PCR and sequencing



Allele specific expression confirms effect of group A variants



Summary

In summary

GWAS is excellent at identifying association

Results are sample size dependent: larger samples → more effects

Chromatin marks help link GWAS results to disease relevant cells

Still working at understanding the *interplay* between different cell types

GWAS alone cannot identify genes

We are getting better at finding causal variants

CHi-C offers a systematic method to link these to genes

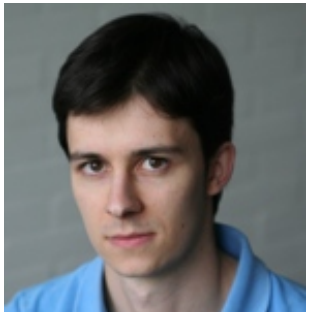
BUT:

CHi-C still noisy, need to access the relevant cell, relevant condition

Why didn't eQTLs work?

Thanks to

Olly Burren



Arcadio Rubio Garcia

Tony Cutler

CHi-C collaboration

Biola Javierre, Jonathon Cairns, Willem Ouwehand, Linda Wicker, John Todd, Mikhail Spivakov, Peter Fraser

Further reading

Javierre et al., Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters *Cell* 2016.