Gaining Biological Insight from GWAS Studies

Lavinia Paternoster

Senior Lecturer in Genetic Epidemiology

MRC Integrative Epidemiology Unit, University of Bristol

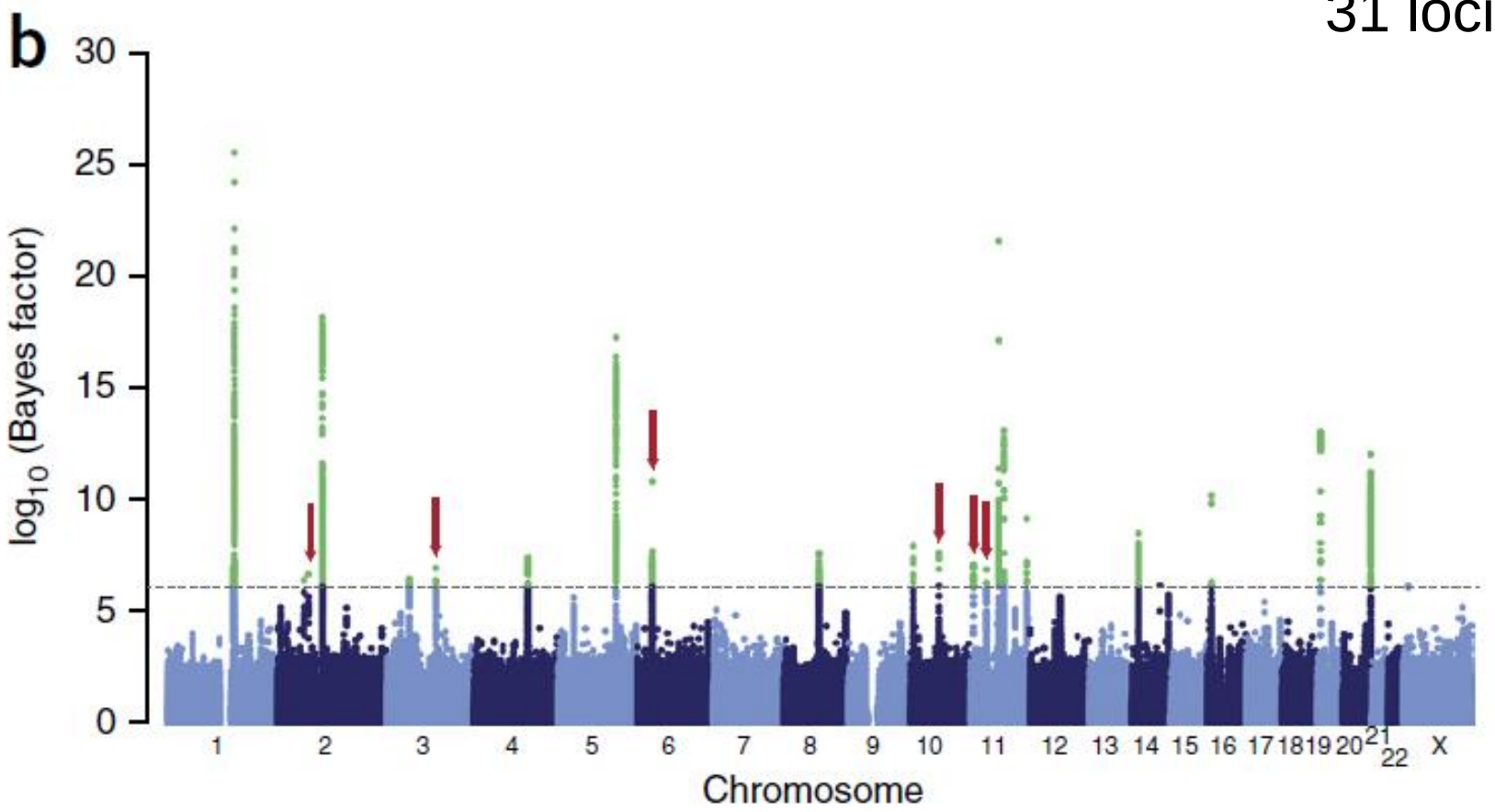
20th November 2018

Outline

- How useful are SNP associations?
- Bioinformatic follow-up
  - The questions
  - Available data & approaches
  - Why it might not work
- Beyond bioinformatics – the next steps
- Your turn!

Reminder: What have AD GWAS found?

31 loci



Paternoster et al. 2015, Nature Genetics 47:1449-56

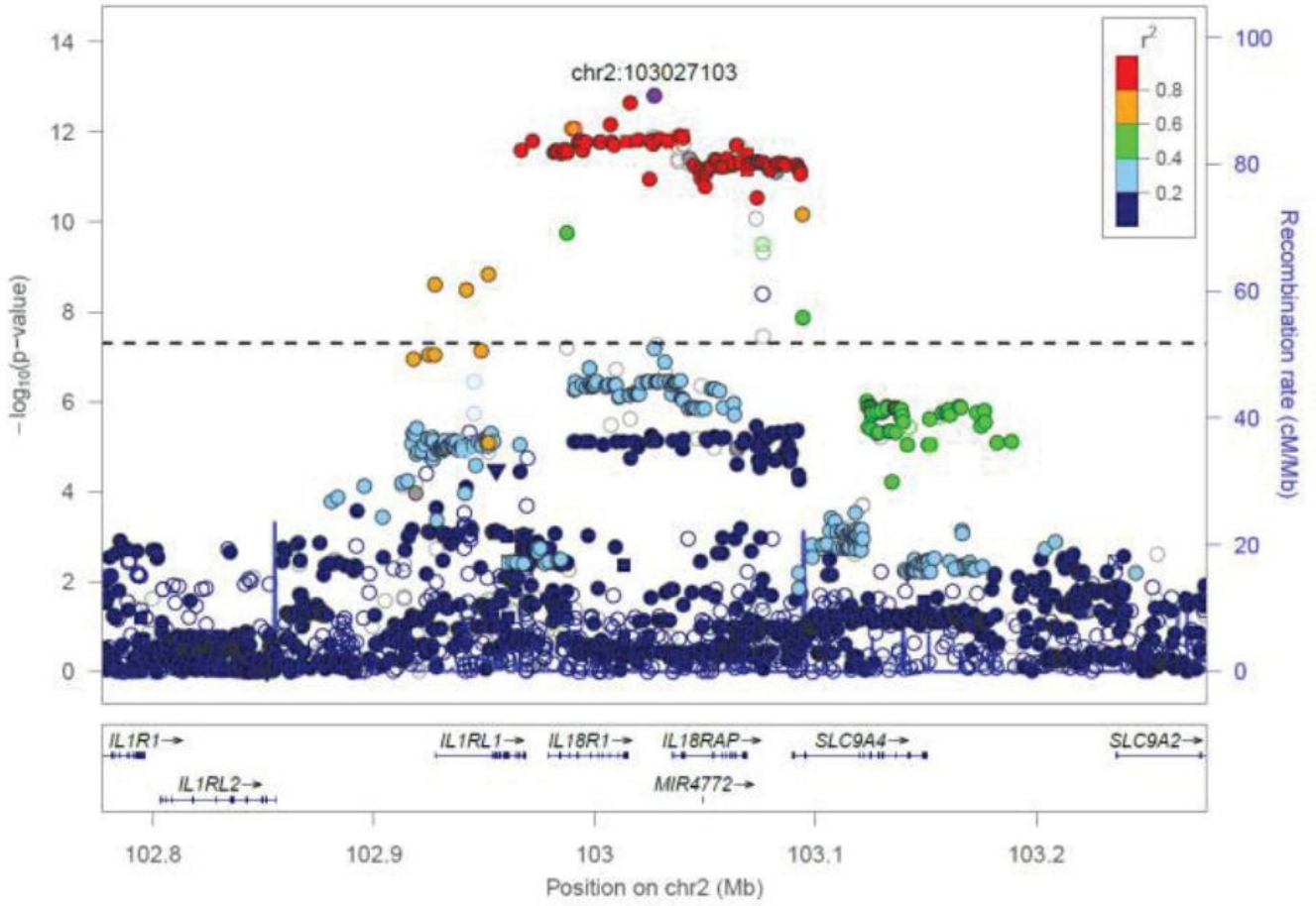
Reminder: What have AD GWAS found?

Variant	Locus	Nearest gene[a]	EA/OA	European, fixed effects				All cohorts, MANTRA		Known atopy loci?	
				n (studies)	EAF	OR (95% CI)	*P*	*n* (studies)	$\log_{10}$ (BF)	Trait	Reference
Known loci											
rs61813875	1q21.3	*CRCT1*/***LCE3E*** (*FLG*)[b]	G/C	93,326 (18)	0.02	1.61 (1.48–1.75)	**5.6×10^{-29}**	96,419 (20)	**25.53**	AD	3,4,5
rs10791824	11q13.1	*OVOL1*	G/A	102,761 (21)	0.57	1.12 (1.09–1.15)	**2.1×10^{-19}**	116,556 (25)	**21.56**	AD	9
rs12188917	5q31.1	*RAD50*/***IL13***	C/T	102,761 (21)	0.21	1.14 (1.10–1.17)	**4.0×10^{-17}**	116,554 (25)	**17.24**	AD, A, IgE	9,18,58
rs6419573	2q12.1	*IL18R1*/***IL18RAP***	T/C	102,760 (21)	0.26	1.11 (1.08–1.14)	**1.5×10^{-13}**	116,557 (25)	**18.10**	AD, A, AS, SRA	8,14,18,21
rs2212434	11q13.5	***C11orf30***/*LRRC32*	T/C	102,761 (21)	0.45	1.09 (1.07–1.12)	**4.6×10^{-13}**	116,557 (25)	**13.02**	AD, AS, SRA, AR, A	11,14,15,21,59
rs4809219	20q13.33	***RTEL1/TNFRSF6B***	C/A	102,760 (21)	0.27	0.90 (0.87–0.93)	**7.0×10^{-13}**	116,555 (25)	**11.98**	AD	7,10
rs2918307	19p13.2	*ADAMTS10*/***ACTL9***	G/A	100,707 (20)	0.16	1.12 (1.08–1.16)	**4.6×10^{-12}**	114,504 (24)	**12.98**	AD	9
rs2041733	16p13.13	*CLEC16A*	C/T	103,066 (22)	0.55	0.92 (0.90–0.94)	**2.5×10^{-11}**	116,862 (26)	**10.11**	AD, A+HF	7,53
rs12730935[c]	1q21.3	*IL6R*	A/G	102,760 (21)	0.39	1.08 (1.05–1.11)	**6.1×10^{-11}**	116,556 (25)	**7.15**	AD, A	12,15
4:123243592[d]	4q27	*KIAA1109* (*IL2*)[b]	R/I	102,761 (21)	0.37	1.08 (1.05–1.10)	**4.2×10^{-9}**	107,119 (24)	**7.32**	AD, AS, SRA	7,14,21
rs4713555	6p21.32	***HLA-DRB1***/*HLA-DQA1*	T/G	91,217 (15)	0.27	0.91 (0.89–0.94)	**5.4×10^{-9}**	105,014 (19)	**10.76**	AD, AS, SRA, A	6,8,14,18,21

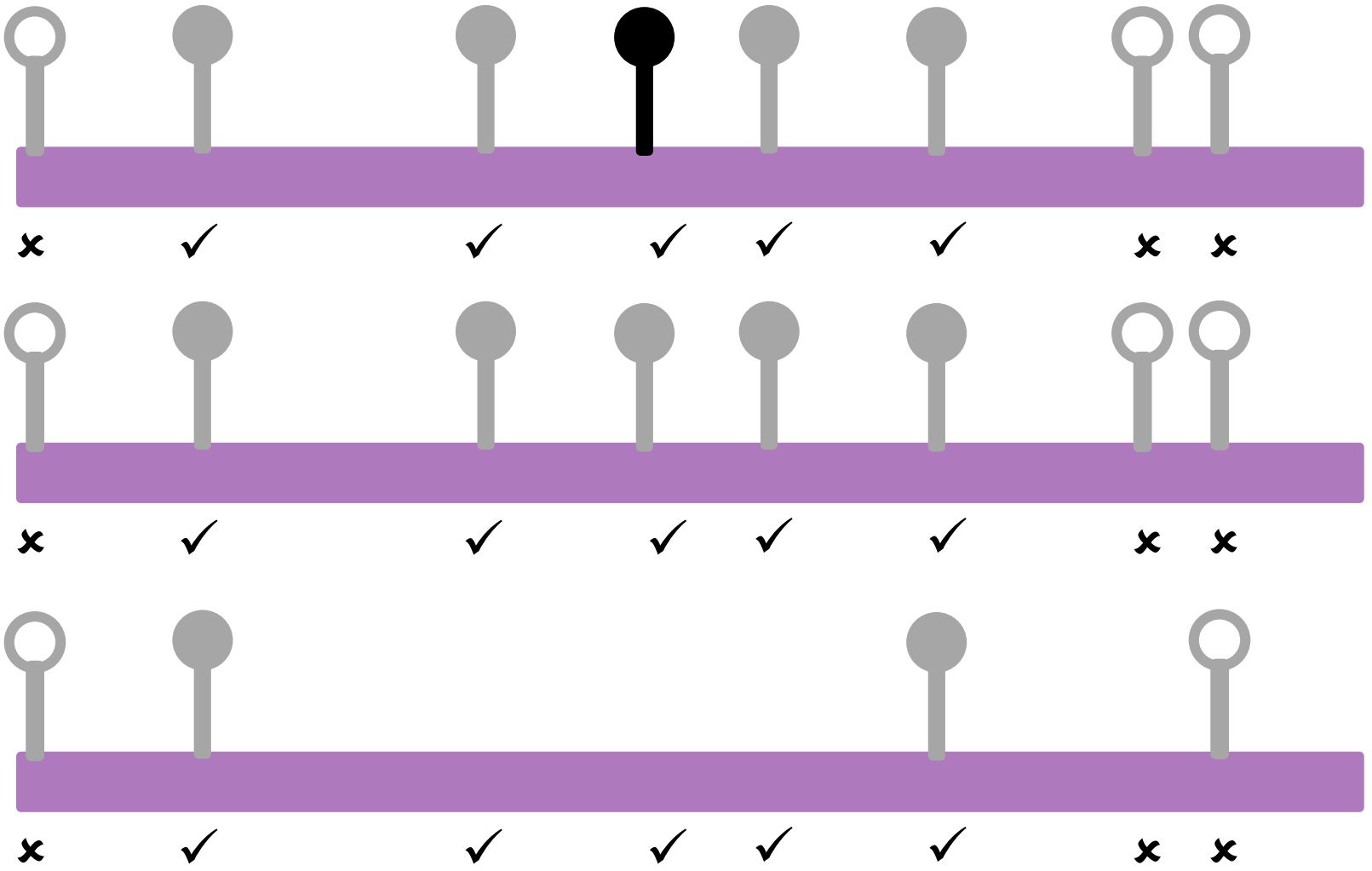
Paternoster et al. 2015, Nature Genetics 47:1449-56

What they haven't found

GWAS identify loci not causal genes or <u>even causal SNPs</u>

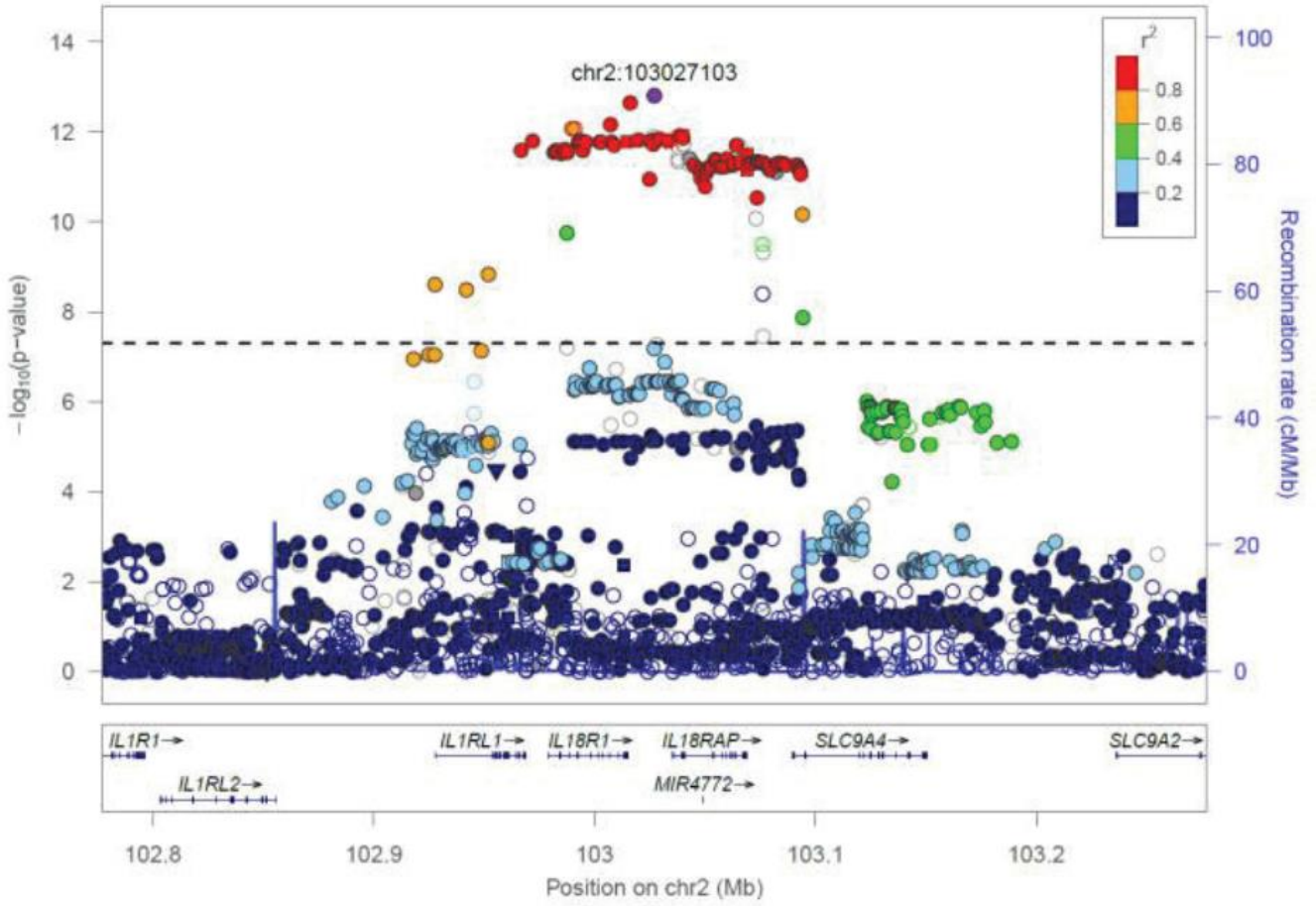


SNPs as markers



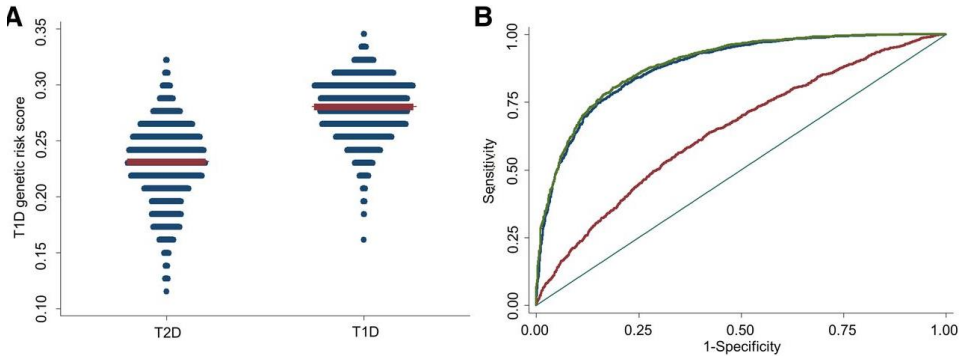
What they haven't found

GWAS identify loci not causal genes or <u>even causal SNPs</u>

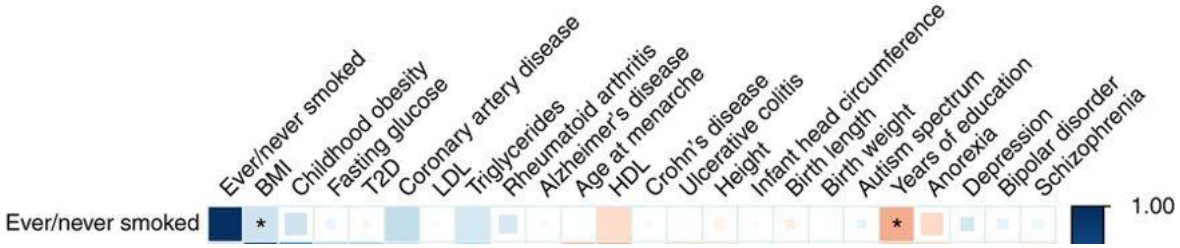


What are genetic marker associations useful for?

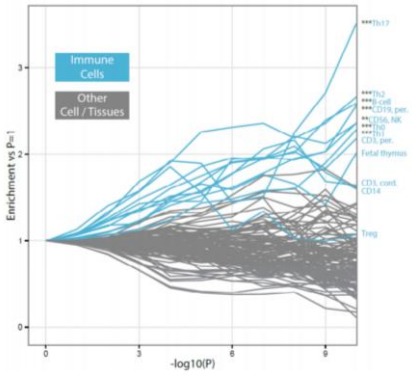
- Prediction



- Estimating genetic correlation between traits

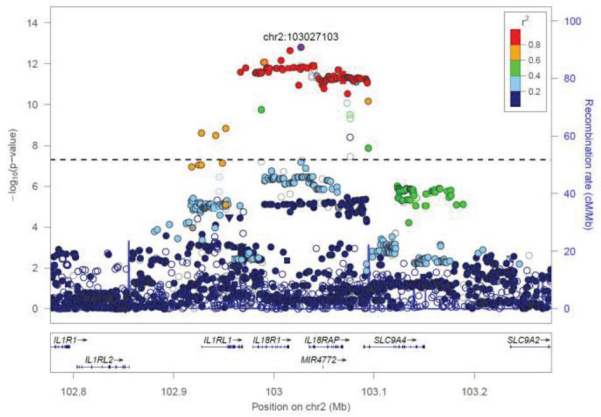


- Enrichment analyses



e.g. what is the overlap between:
- Disease associated regions &
- Regions in open chromatin (across many cell types)

Understanding biology at individual loci



The questions:

- What is the causal variant?
- What is the causal gene?
- Which cell type is affected?
- What is the mechanism/pathway that is disturbed?

What is the causal variant?

Fine-mapping

- Statistical methods that use relative effect sizes to determine a **credible set** (e.g. 95% credible set)

  Software: FINEMAP, PAINTOR, CAVIARBF, JAM

Explore variant functions

- Any coding variants in credible set?
- Examine predictions of functional consequences

  Online tool: Variant Effect Predictor

- Explore experimental evidence that variants have regulatory function (e.g DHS, methylation marks etc. from ENCODE, ROADMAP)

  Online browsers: UCSC or ensembl

  Online tools: Haploreg or regulomeDB

HaploReg v4.1

What is the causal gene?

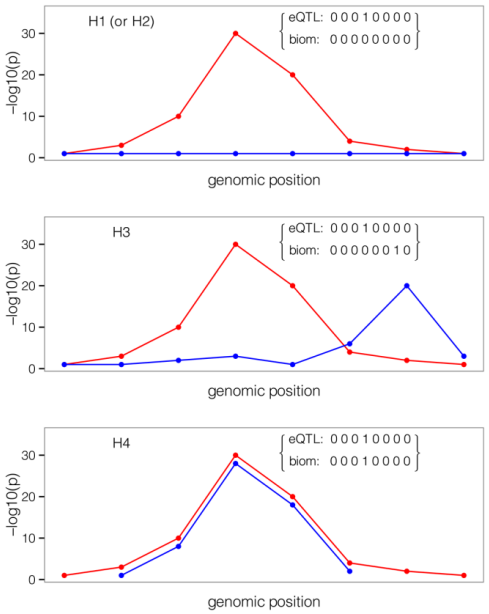
Are disease-associated SNPs also expression-associated SNPs?

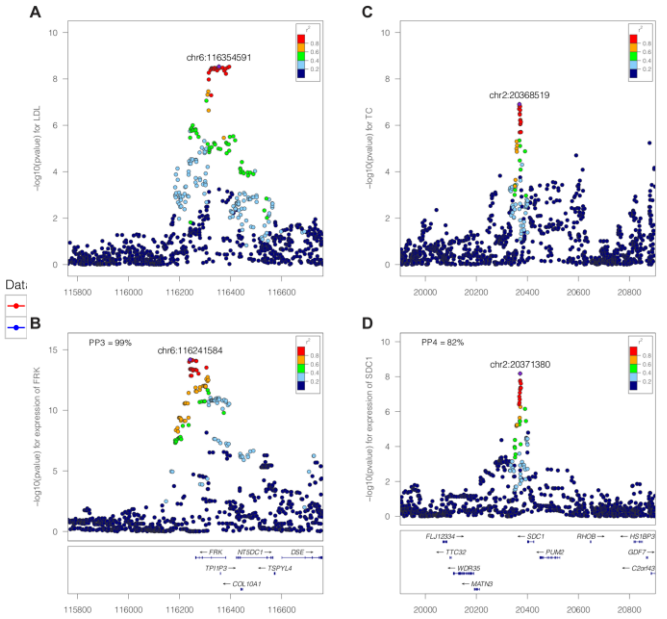
- Explore expression quantitative trait loci (eQTL) results

  (e.g. GTEx, eQTLgen)

- Where full eQTL summary statistics available, conduct Colocalisation analysis

  (R package: coloc, Giambartolomei, PLoS Genet, 2014)

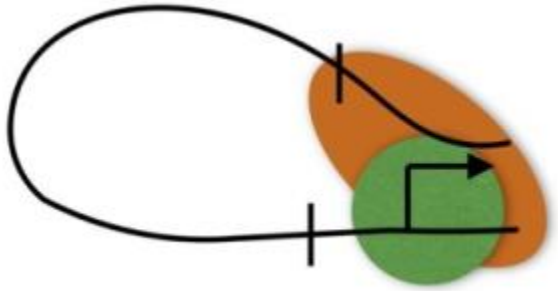




What is the causal gene?

Do disease-associated regions interact in 3D chromatin with gene promoters?

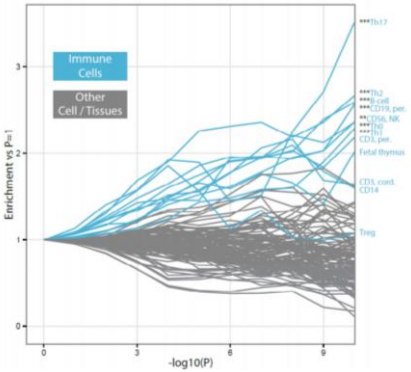
- Use chromatin conformation capture (3C) data to identify which genes an enhancer region acts upon (3C, 5C, Hi-C)



- Online tool: 3D Genome Browser (YUE lab, Penn state):
  promoter.bx.psu.edu/hi-c

Which cell type is affected?

- Genome-enrichment analysis suggest **overall important cell types**, by comparing disease-risk loci with:
  - DHS
  - Or any other cell-specific annotation

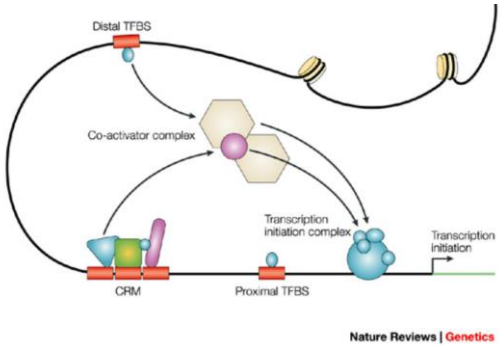


- **At individual loci**:
  - Most annotations I've mentioned are cell-type specific (e.g. expression, DHS, methylation marks, chromatin interaction)
  - So ----- always consider and compare all/relevant cell types

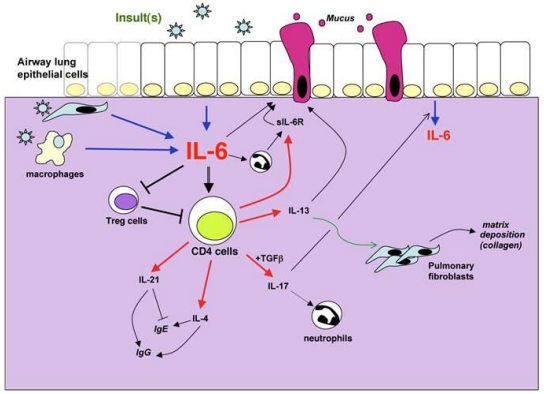
- **BUT, there's an issue:**
  - Requires cell type of interest to be available in public data

What is the mechanism/pathway that is disrupted?

- **Mechanism**
  - Build evidence through the various annotations:
    e.g. If causal SNP is in an enhancer
    – does it disrupt transcription factor binding?



- **Pathway**
  - Which biological pathway does the causal gene sit in?
  - Are other genes in this pathway also implicated by GWAS?
  - If designing new drugs, where might be the best point to intervene?
  - Tools: e.g. Genecards, MAGMA



Integrated online tools

- Genome browsers:
  - UCSC
  - Ensembl
  - Both allow you to add relevant data ‘tracks’

- Regulation summaries:
  - Haploreg
  - regulomeDB
  - FUMA

- Simple SNP -> Gene tool:
  - OpenTargets - Genetics

If you only do one thing……..

(my opinion)

Compare GWAS results with eQTL data from relevant tissues

(But only because eQTL data is currently much richer than pQTL data)

Why it might not work……

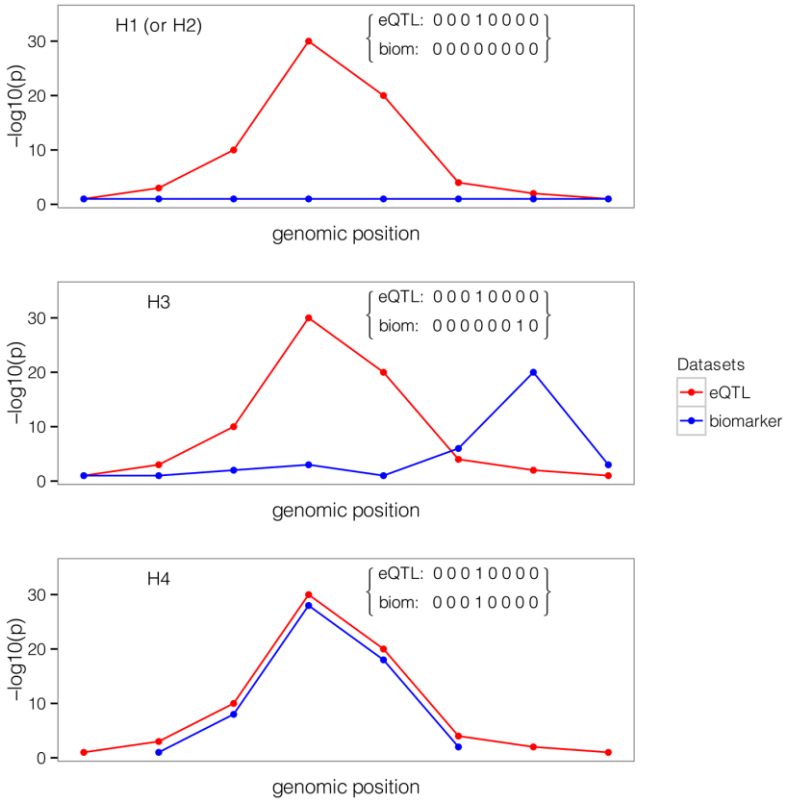
- In silico bioinformatic analyses are only as good as the data they use!
  - Important **cell type** may not have been studied
  - Effect might only be seen under certain **conditions**
  - Effect might only be seen in **diseased samples**
  - Available **sample sizes** might be insufficient

- Also beware **Red Herrings**:
  - Just because a GWAS region is annotated as an enhancer in LCLs, it doesn't automatically follow that this is relevant!

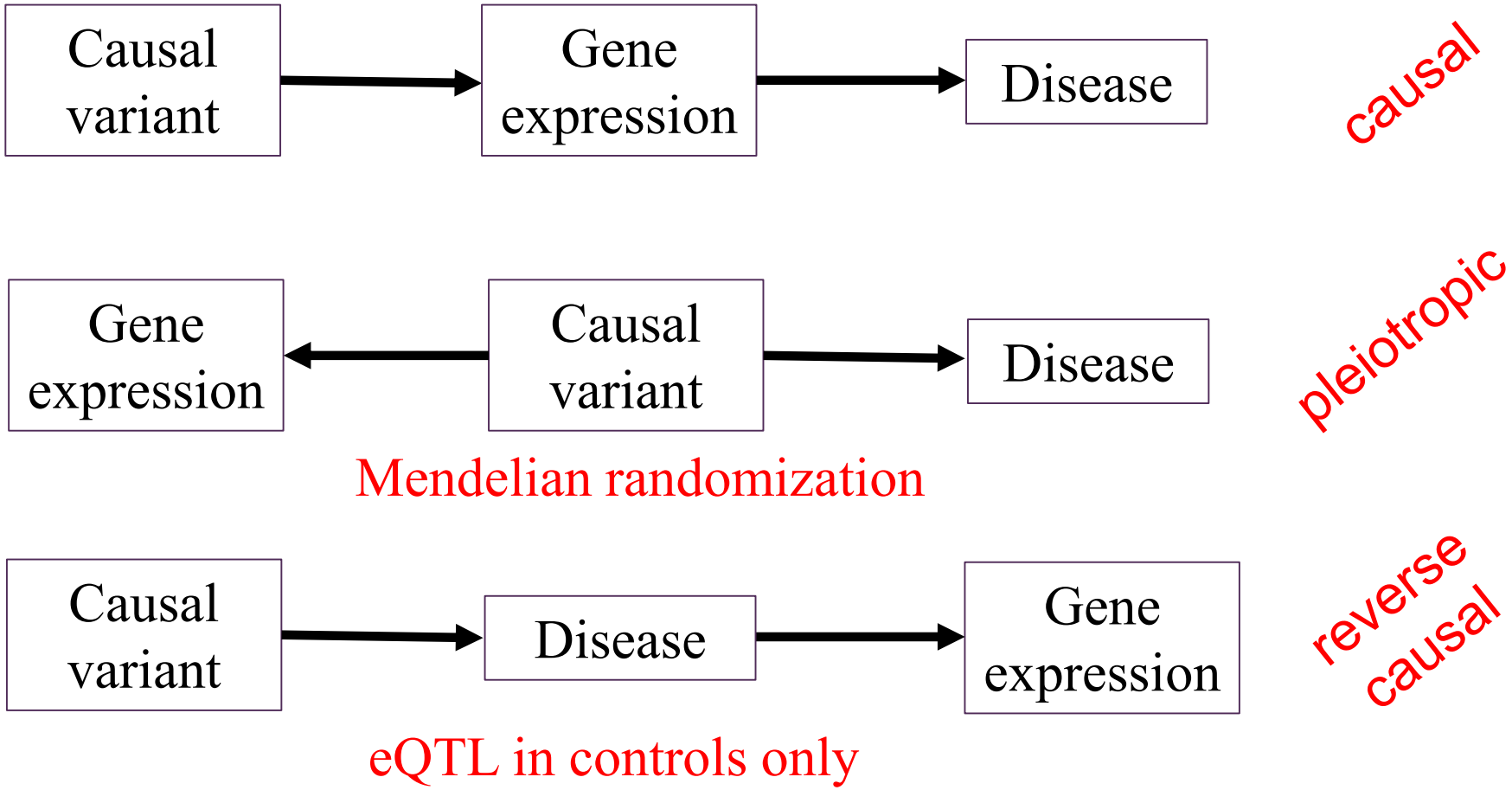
Correlation ≠ Causation!

One last problem……

Does colocalization infer causality?

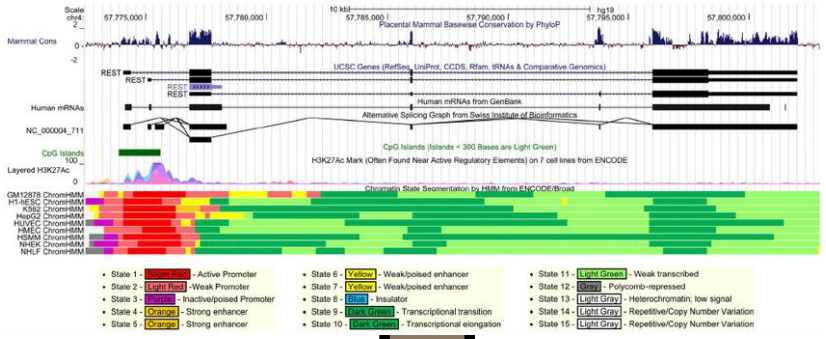


Possible colocalization mechanisms

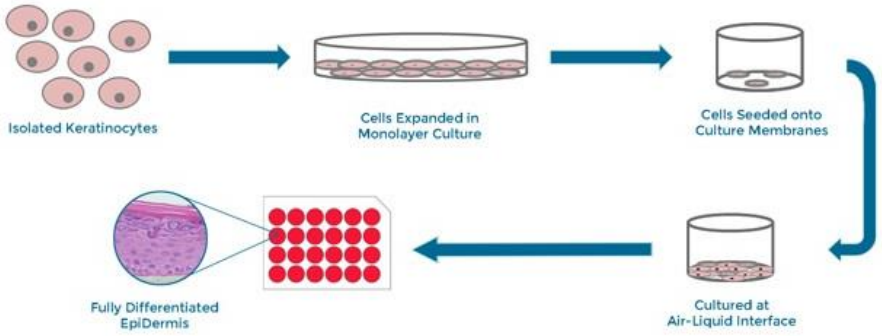


In vitro/in vitro experiments

Build hypotheses with in silico bioinformatic data

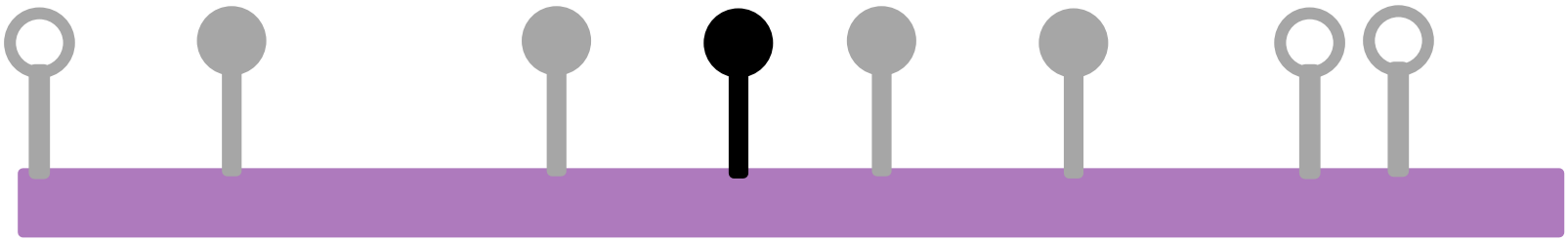


Test these hypotheses with carefully designed in vitro/in vivo experiments



Causal genes/cell-types/variant

- If you only care about identifying the gene (& cell type) you can to some extent not worry about the causal variant
- BUT……if you want to do CRISPR experiments to confirm the causal gene, you'll need to edit the right variant



Practical instructions

Navigate to: github.com/epxlp/ADbioinformatics

Follow ‘instructions.md’

Main aims of this practical:

- Try out some of the tools I have mentioned (make a note of the ones you prefer or find most useful)
- Attempt to identify causal gene and/or cell types for an AD GWAS locus

- Think about what we haven’t done in this practical that you would want to do

Take home messages

- There is a lot of useful publicly available information out there
- But, trying to properly interpret it is very difficult

DO

- When you read papers that use such investigations to attempt to identify causal mechanisms
  - remember the limitations
  - question the findings

READ

- Remember:
  - Is the most relevant data even available? (e.g. cell type)
  - Findings should be validated in experiments

DO

READ