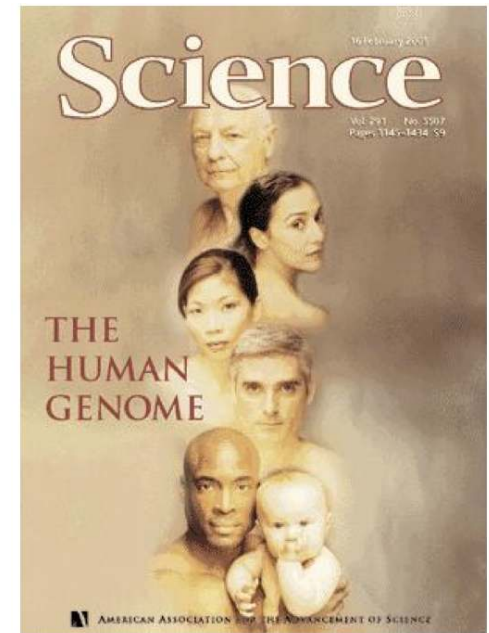


Using epigenetic marks for biological interpretation of complex trait associations

Understanding disease biology

- Using SNPs to map disease loci for complex traits
 - Immune-related diseases, coeliac disease
- From associated variants to function
 - Correlation with gene expression
 - Causal variants, tissues and pathways

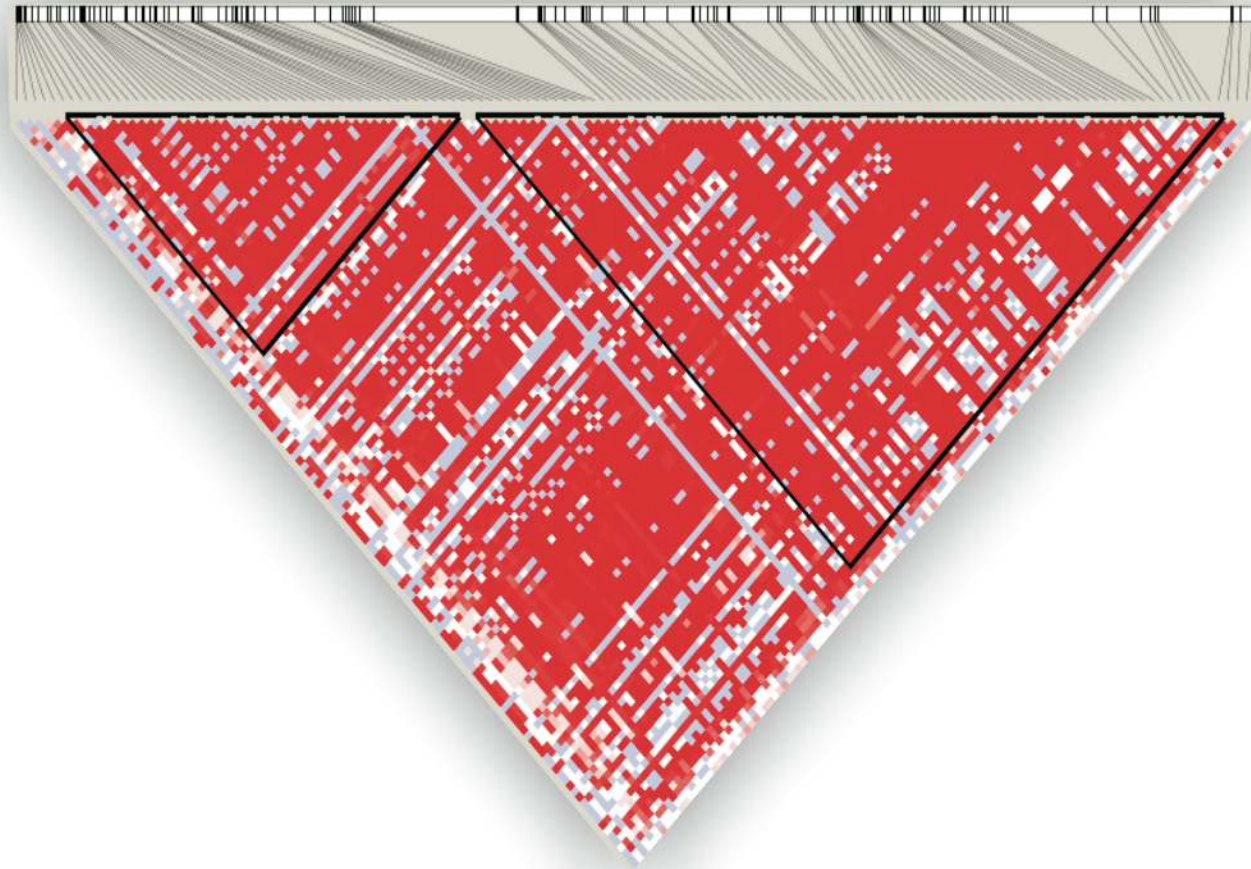
- Sequencing of the human genome
- HapMap project
- 1000 Genomes Project



Reference panels of human variation: HapMap

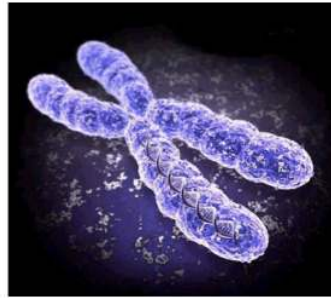
- Goal: **provide a haplotype map of human genome using common variants**
 - Genotyped 1.6 million common SNPs
 - 1,184 reference individuals
 - from 11 global populations
 - Northern and western Europe (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); Yoruba in Ibadan, Nigeria (YRI); African ancestry in the southwestern USA (ASW); Chinese in metropolitan Denver, Colorado, USA (CHD); Gujarati Indians in Houston, Texas, USA (GIH); Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); Mexican ancestry in Los Angeles, California, USA (MXL); Tuscans in Italy (TSI)

Correlation of genetic variants - LD blocks

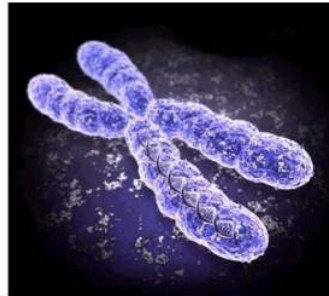


LD – your best friend, your greatest enemy

3 billion base pairs in the human genome

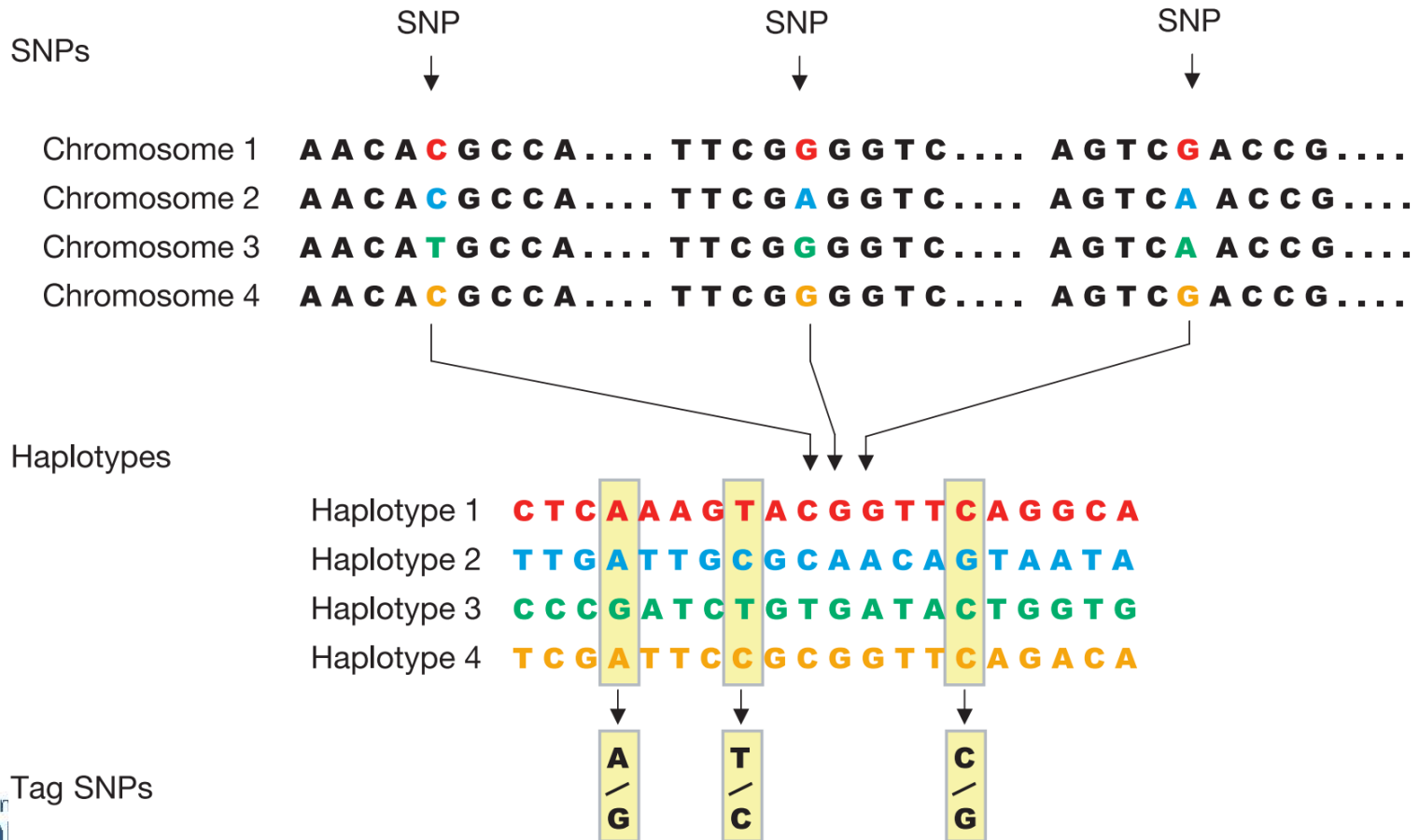
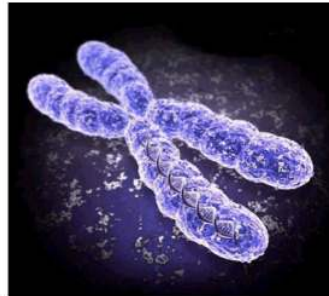


4 million common



SNPs	SNP	SNP	SNP
	↓	↓	↓
Chromosome 1	A A C A C G C C A T T C G G G G T C A G T C G A C C G		
Chromosome 2	A A C A C G C C A T T C G A G G T C A G T C A A C C G		
Chromosome 3	A A C A T G C C A T T C G G G G T C A G T C A A C C G		
Chromosome 4	A A C A C G C C A T T C G G G G T C A G T C G A C C G		

500k tagging SNPs



Imputing missing genotypes

Pin it

Typical imputation scenario

HapMap or
1,000 Genomes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Reference
haplotypes

Cases and
controls typed
on SNP chip

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

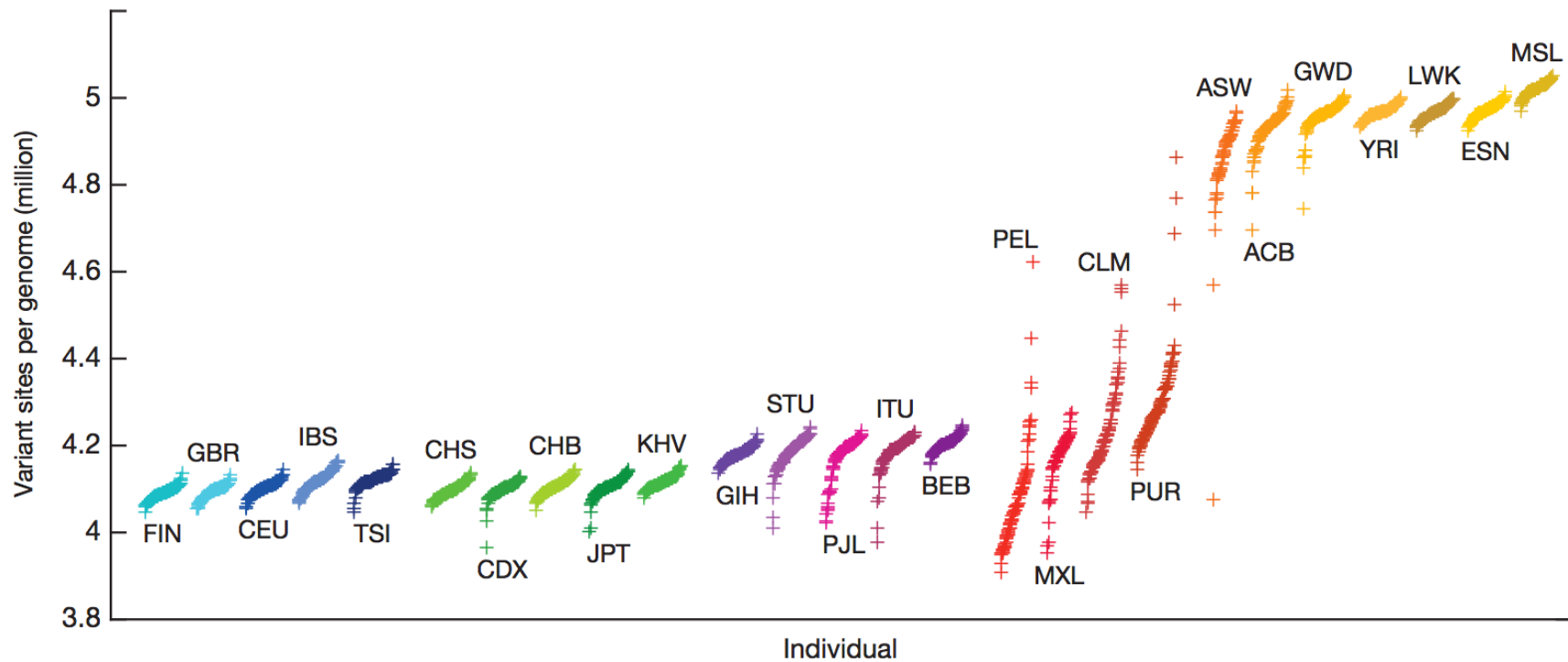
Study
genotypes

Reference panels of human variation: 1000 Genomes Project

- 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping
- > 88 million SNPs, 3.6 million indels, and 60,000 structural variants
- phased onto high-quality haplotypes
- >99% of SNP variants with a frequency of >1% for a variety of ancestries

The total number of observed non-reference sites differs greatly among populations

b



Putatively functional variation

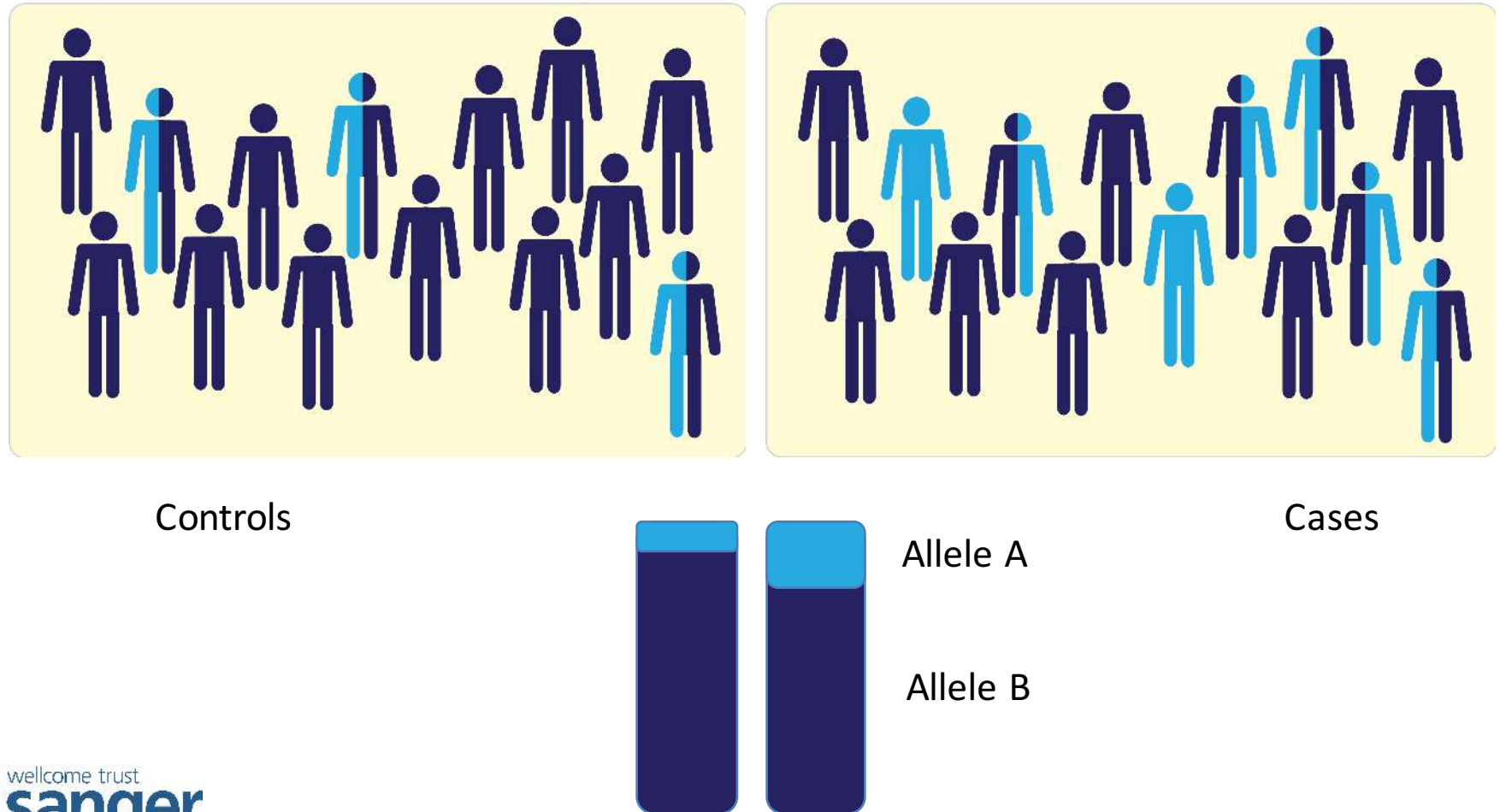
- The majority of variants in the data set are rare but the majority of variants observed in a single genome are common
- A typical genome contained 149–182 sites with protein truncating variants
- ~2,000 variants per genome associated with complex traits through genome-wide association studies (GWAS)
- 24–30 variants per genome implicated in rare disease through ClinVar

Haplotype Reference Consortium

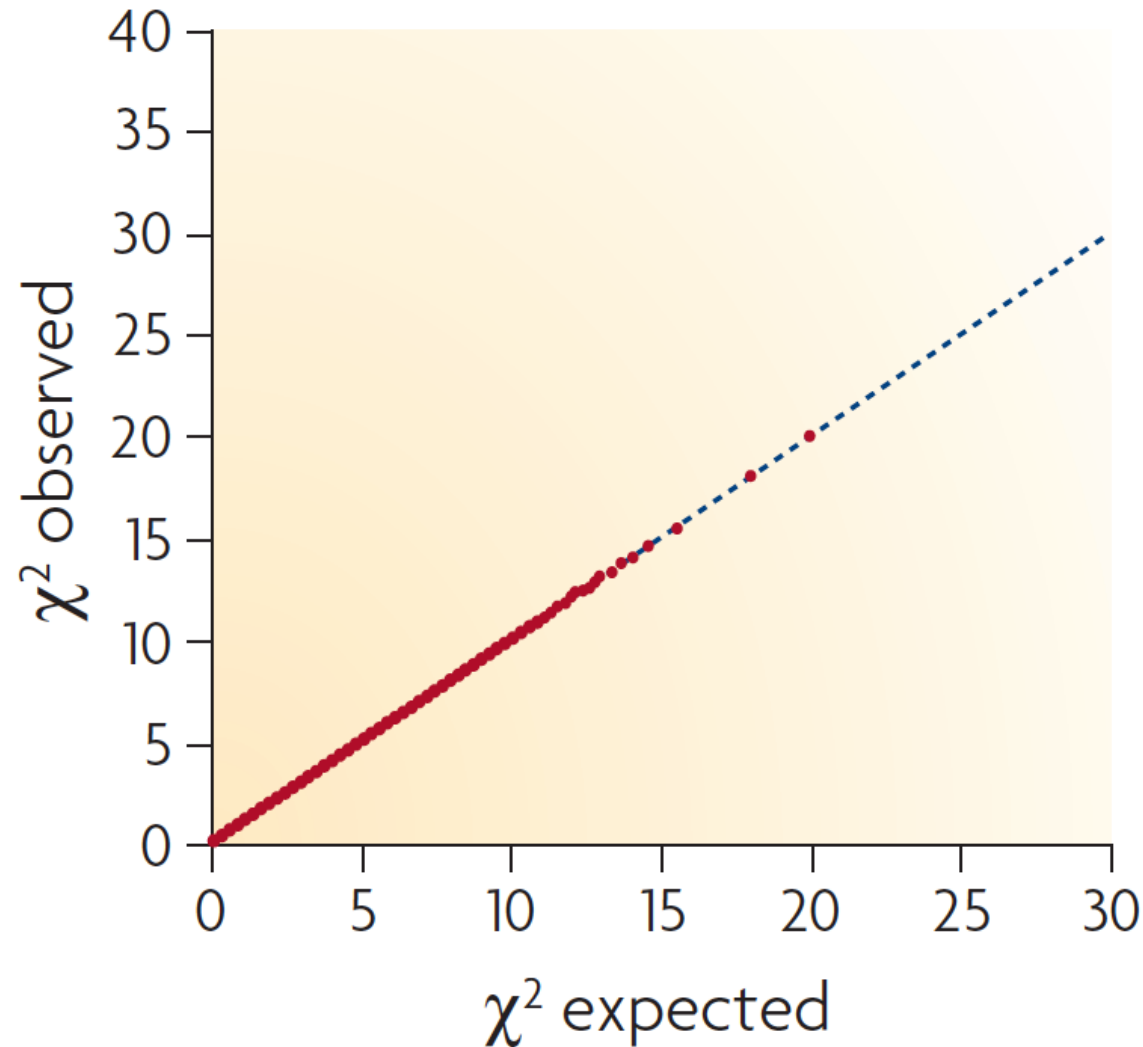
- <http://www.haplotype-reference-consortium.org>
- The first release consists of 64,976 haplotypes at 39,235,157 SNPs, all with an estimated minor allele count of ≥ 5

Genome wide association study (GWAS)

- GWAS compares allele frequency at common SNPs between cohorts of cases and controls

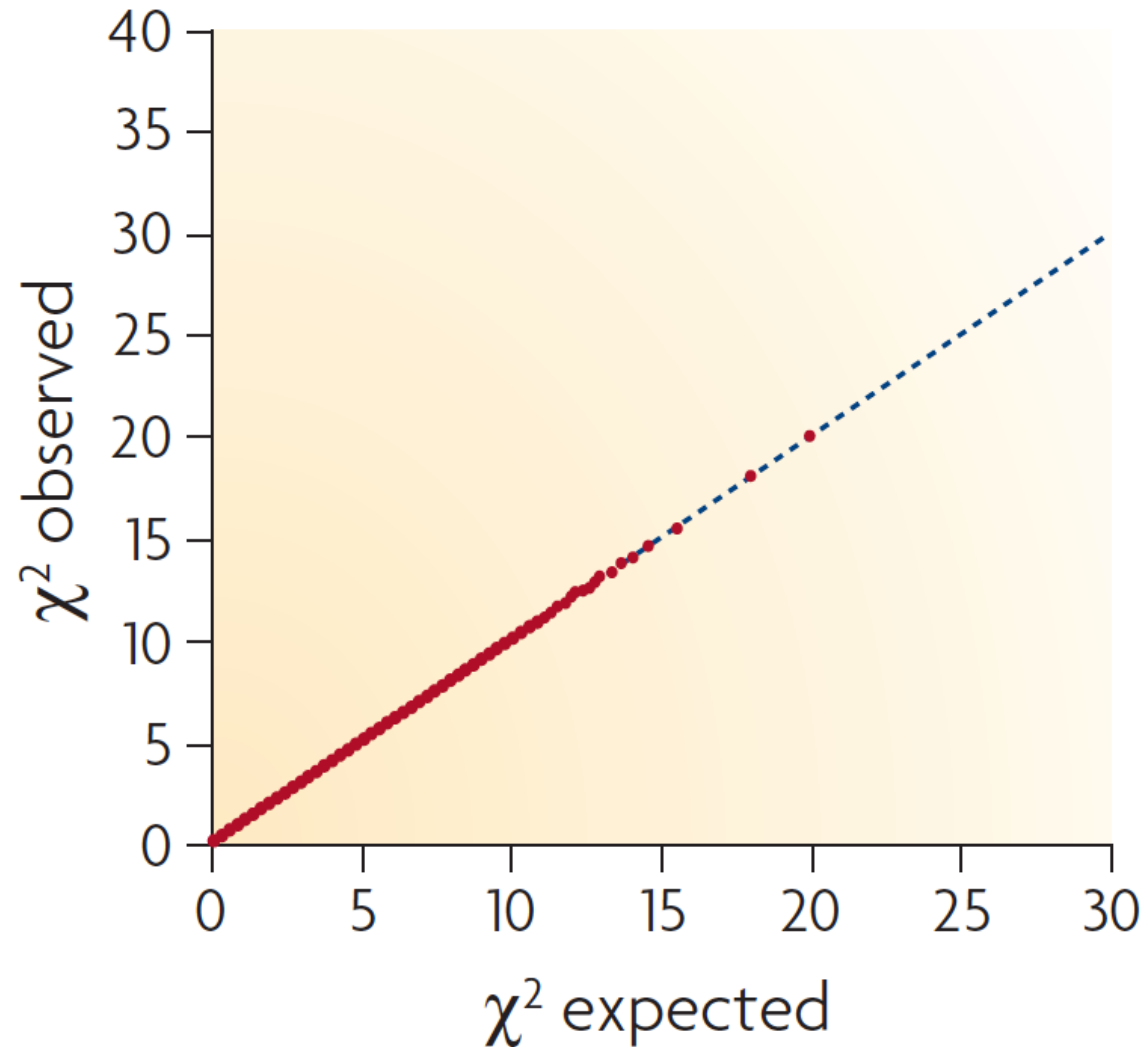


Quantile-quantile (Q-Q) plots



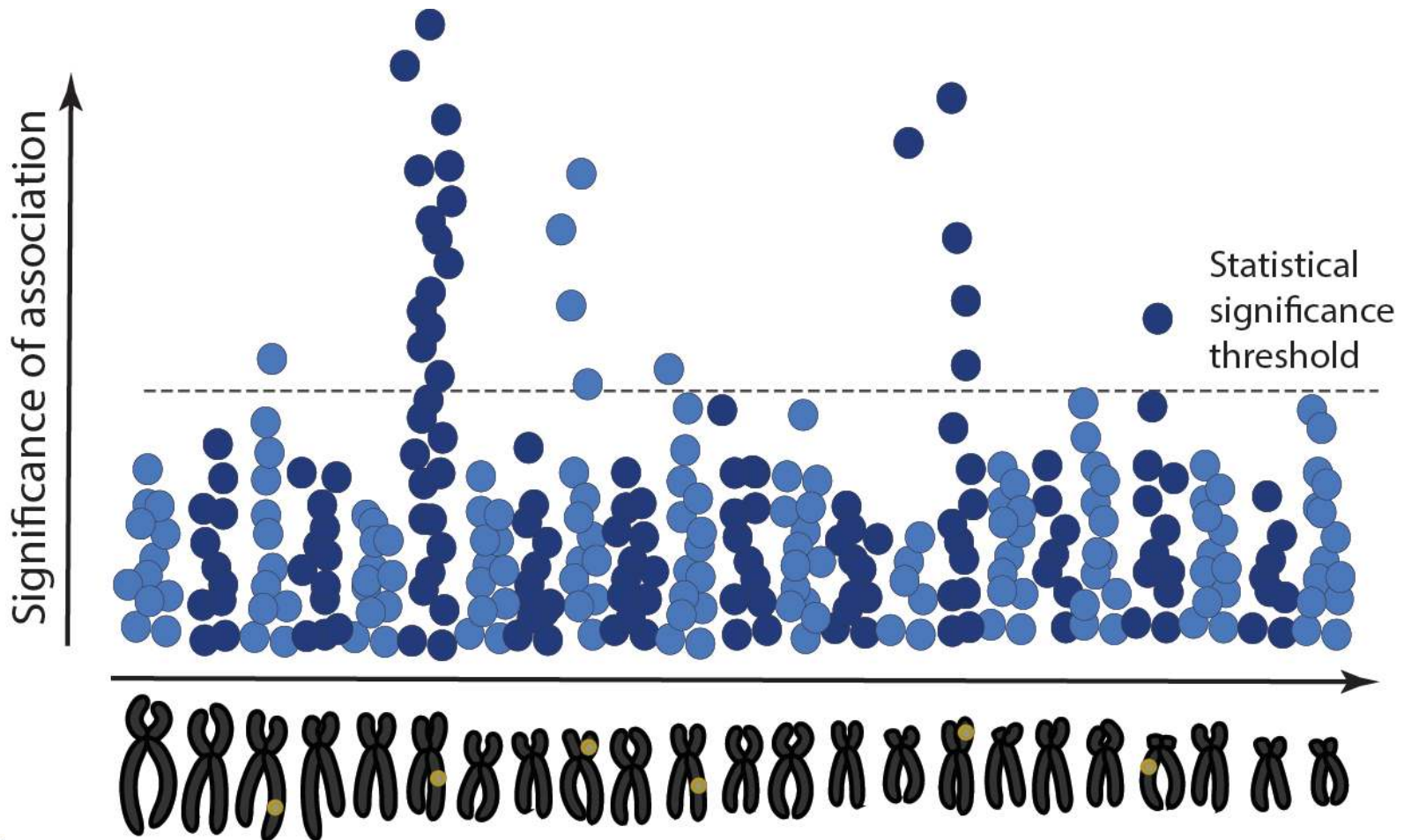
McCarthy et al., *Nat Rev Genet* 2008

Quantile-quantile (Q-Q) plots

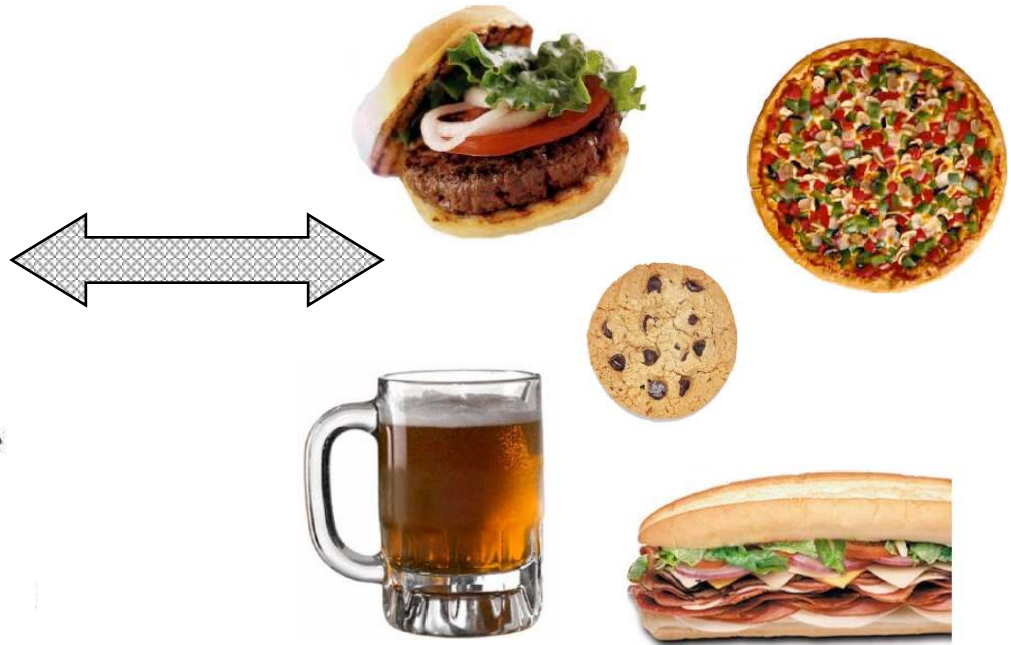
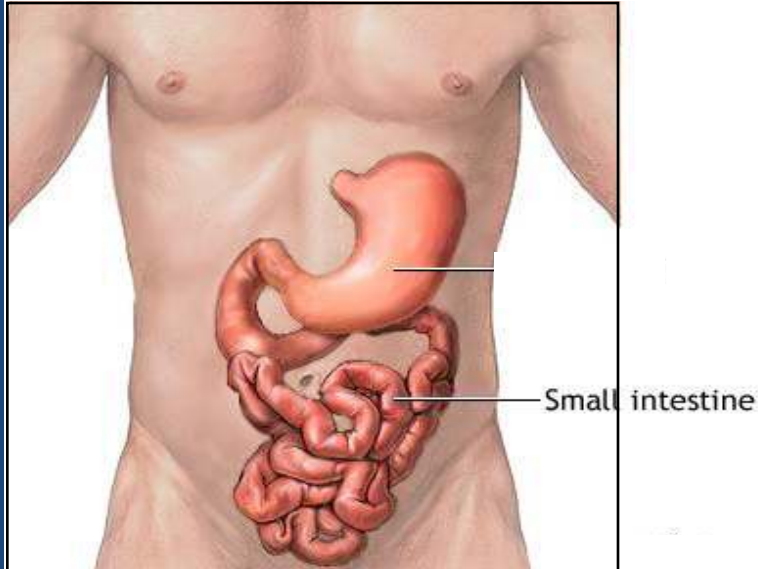


McCarthy et al., *Nat Rev Genet* 2008

‘Manhattan’ plot represents significantly associated regions in the genome



Celiac disease – autoimmune disease of the small intestine

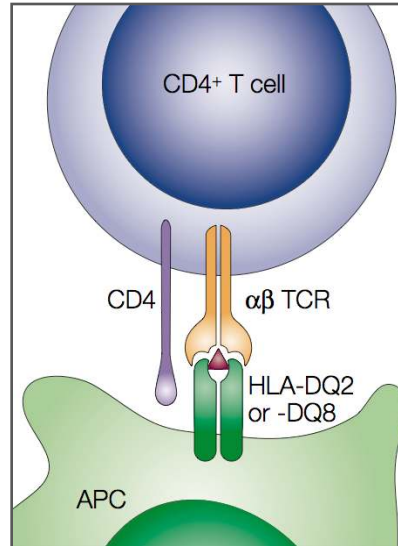


The most common food intolerance in western populations (**~1%**)

Unique understanding of pathogenesis of coeliac disease



+



+

More
genes?

=

Celiac
disease

Wheat, Barley,
Rye

HLA DQ2/DQ8

HLA accounts for about 35% of the genetic variance

The same genotypes are present in 30-40% of general population

Mapping risk variants for coeliac disease

nature
genetics

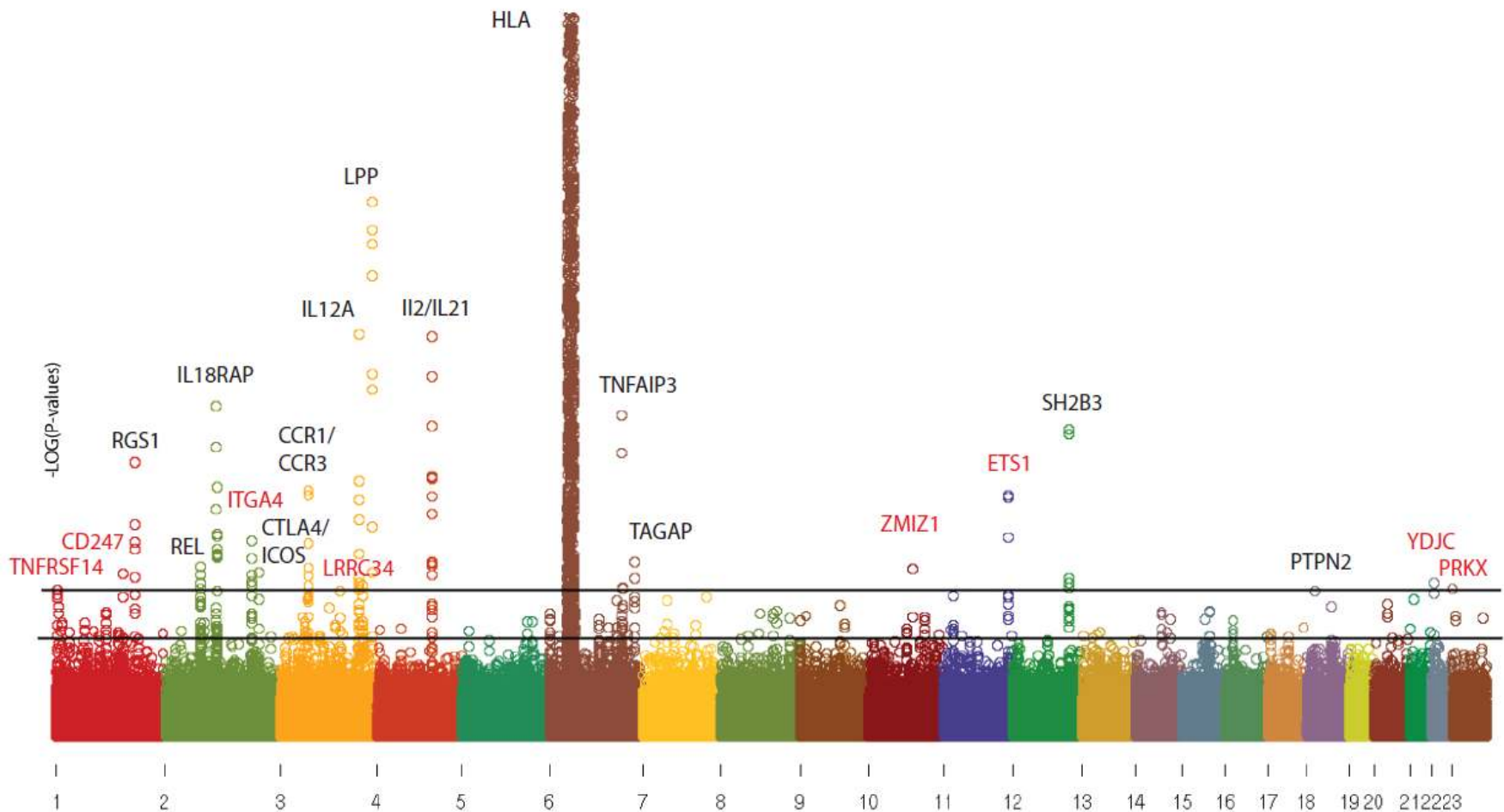
- 800 cases and 1,400 controls from UK

A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*

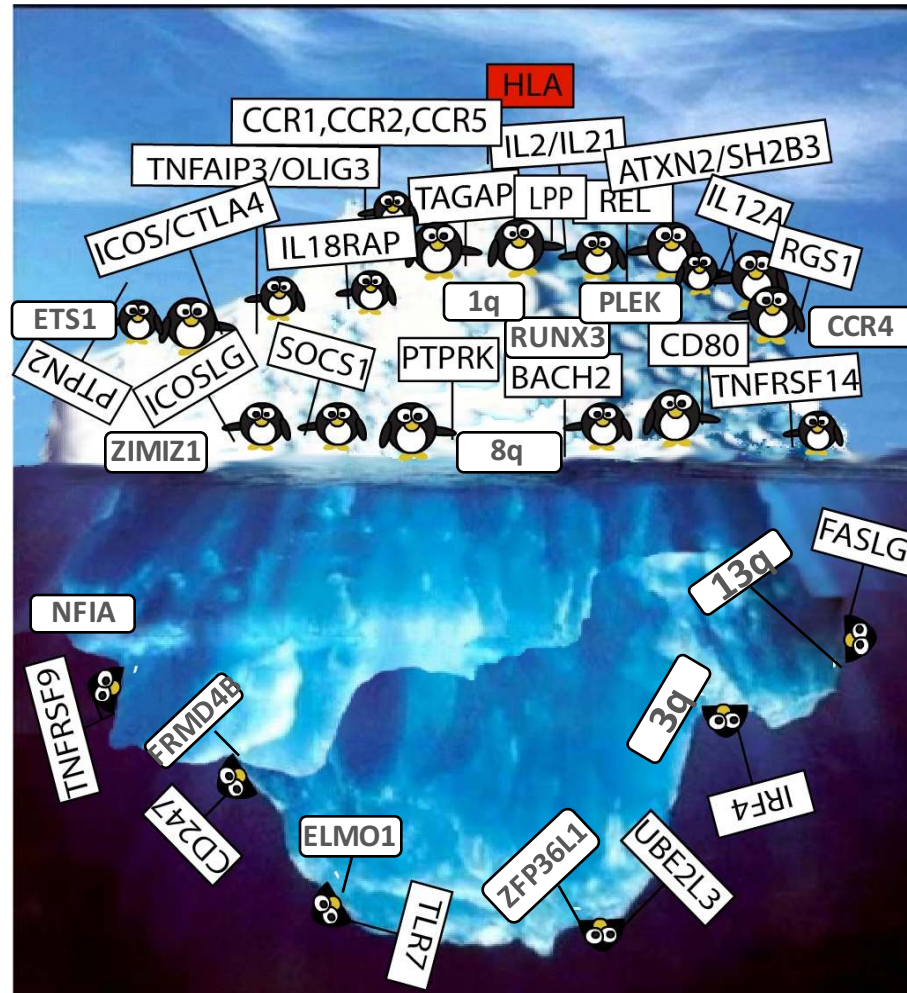
Van Heel, *Nat Genet* 2007

- Increased power with increased sample size and denser SNP genotyping
 - 4,533 cases and 10,750 controls from four populations
 - Top 131 SNPs genotyped in a replication cohort (10,602 samples, 7 populations)

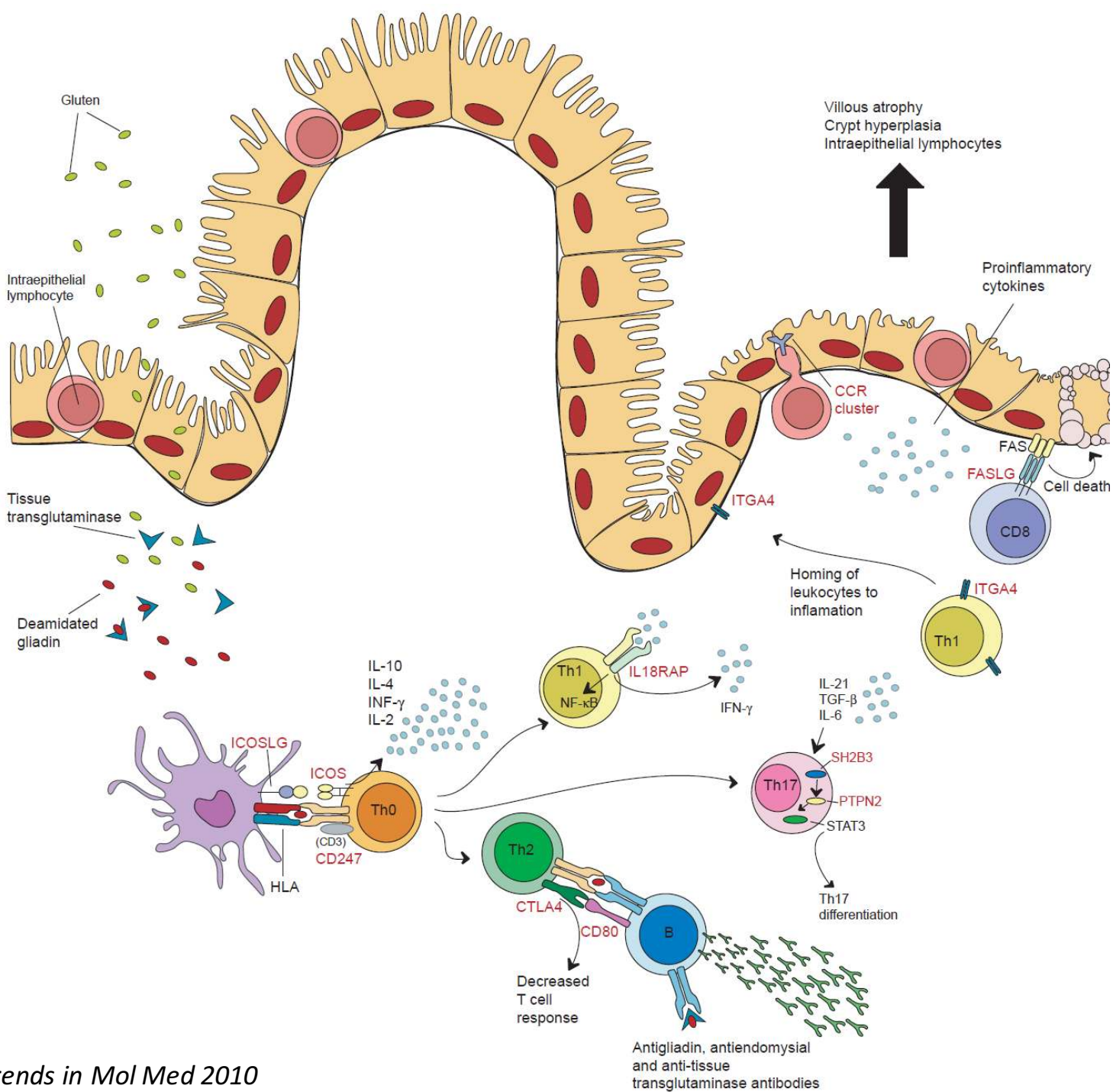
GWAS on 10k cases and 16k controls from 10 populations – 26 associated regions



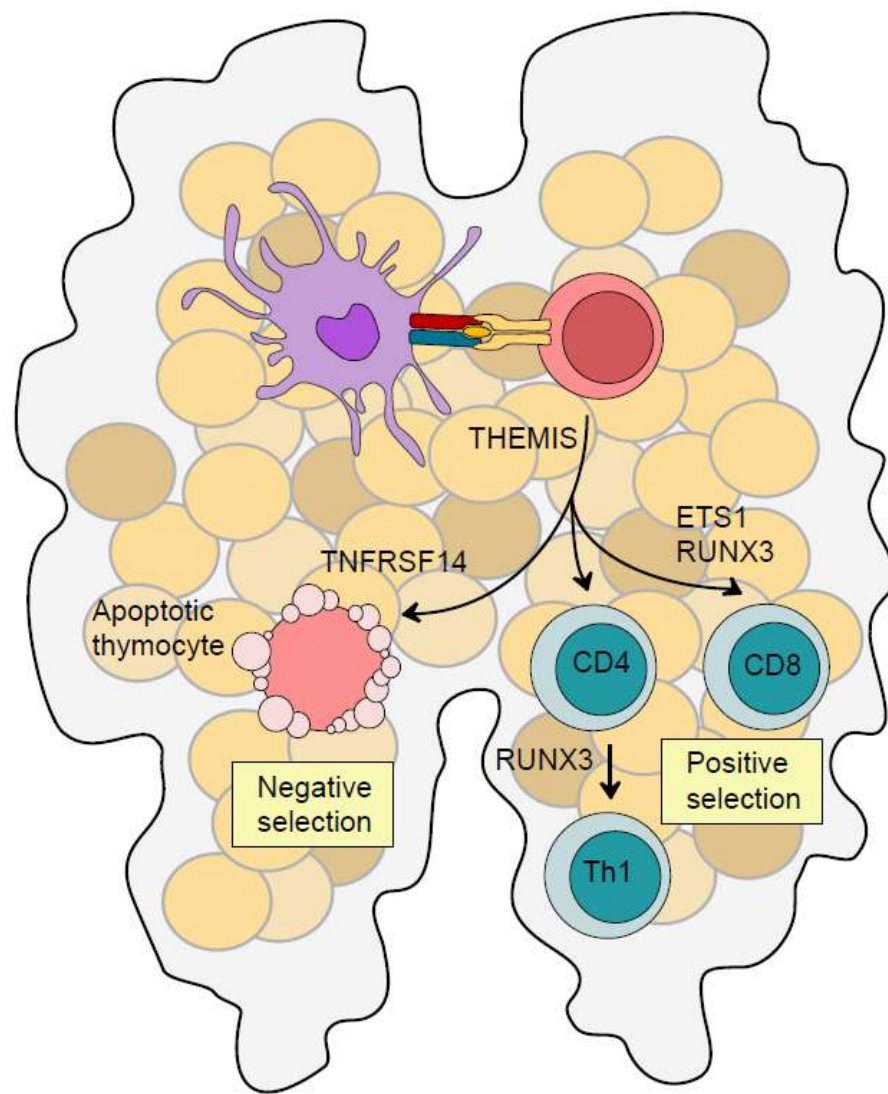
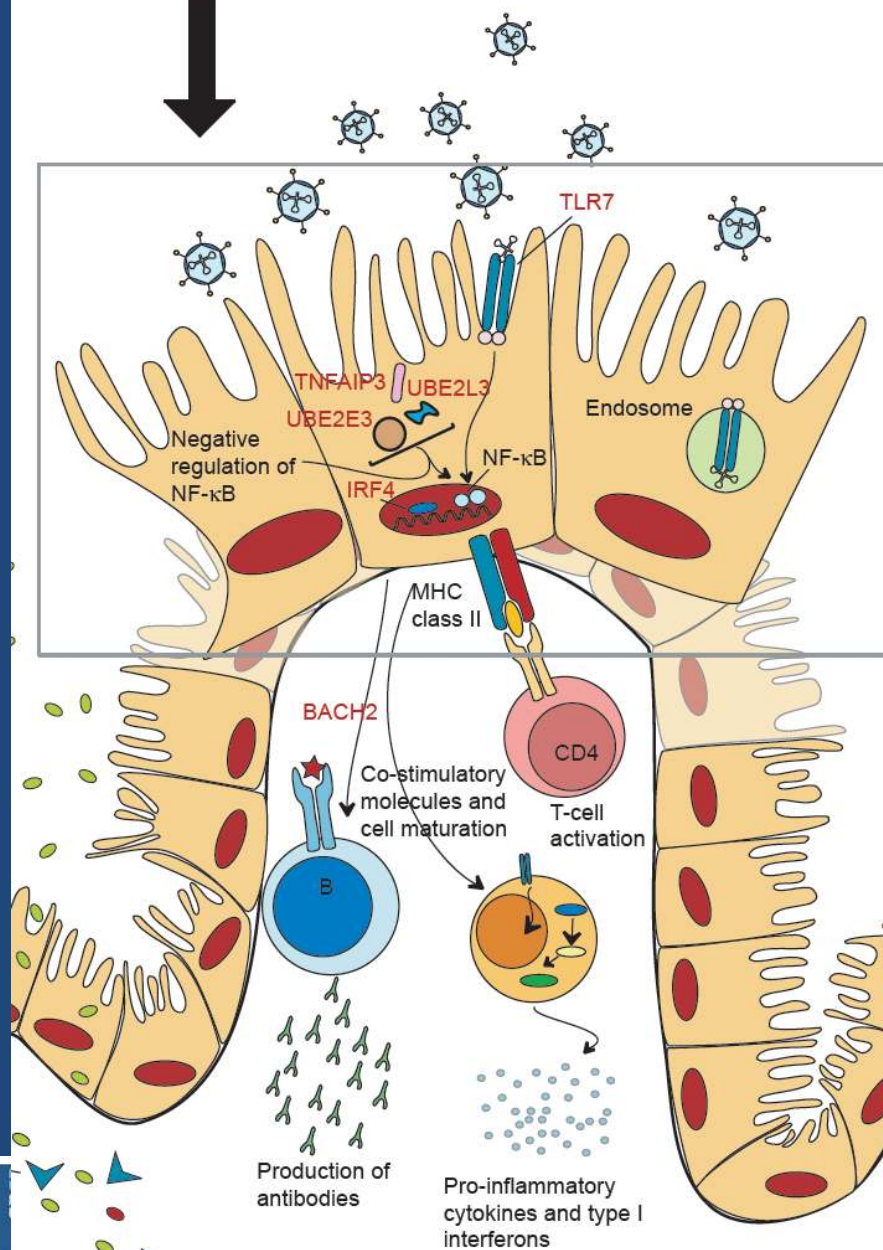
HLA and 26 loci explain 40% of genetic risk



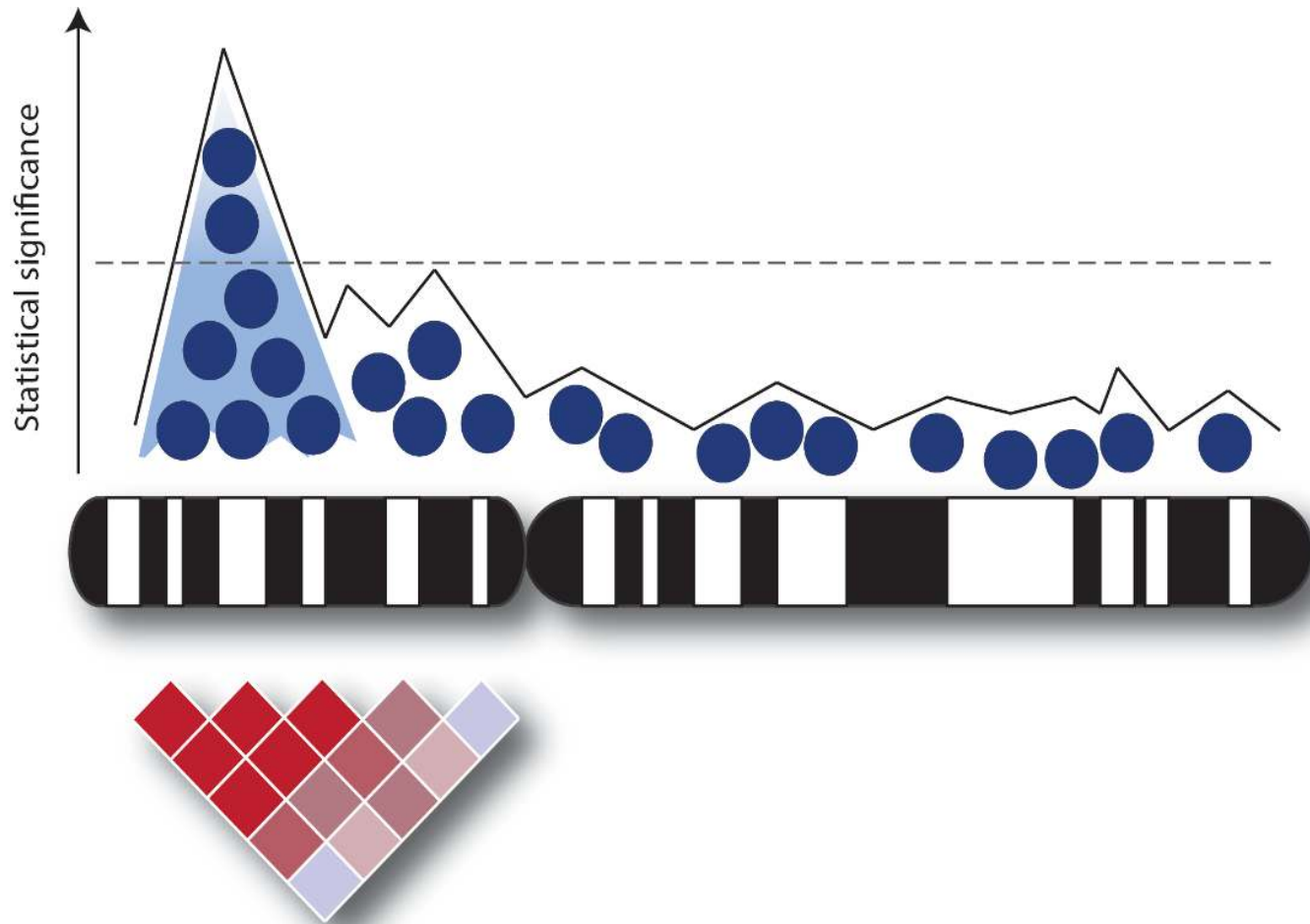
- Genes in proximity of associated variants confirm known and implicate novel pathways



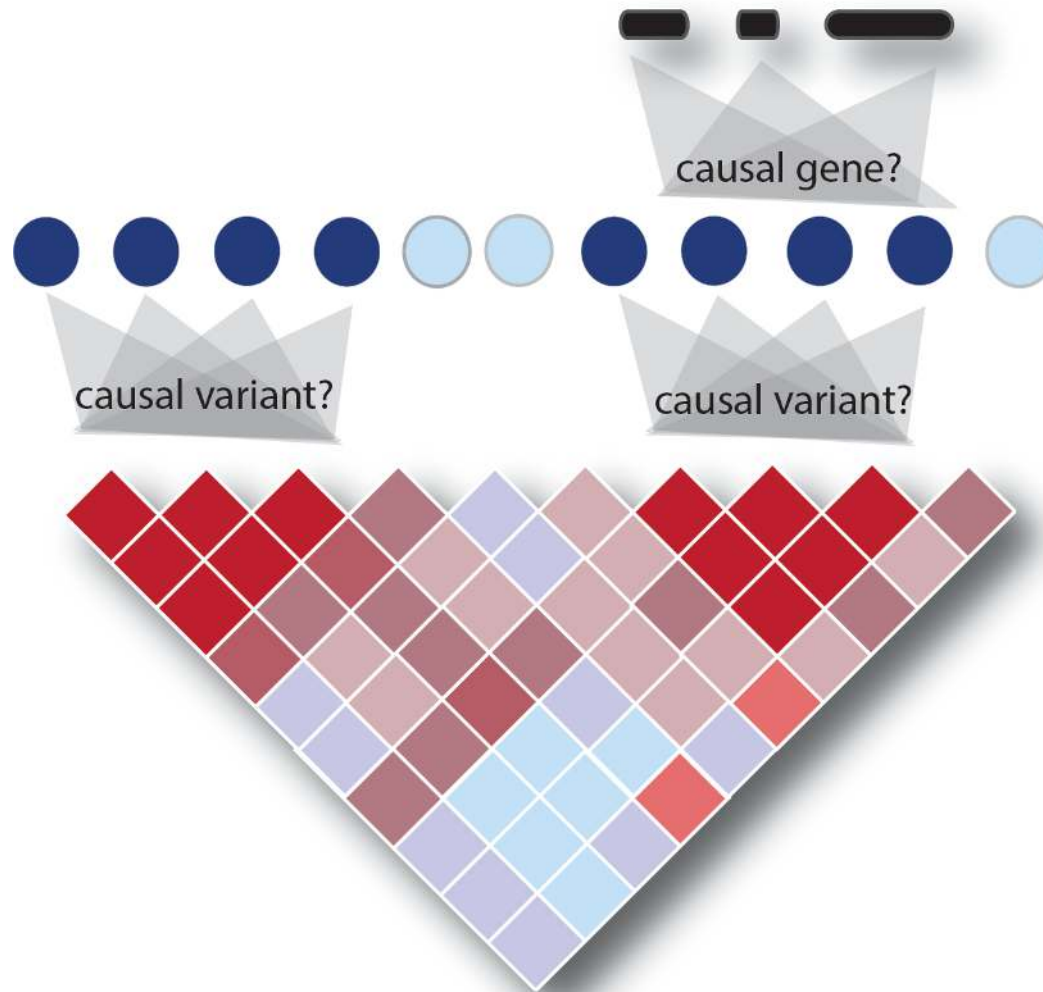
Environmental factors
Bacterial/virus infection?



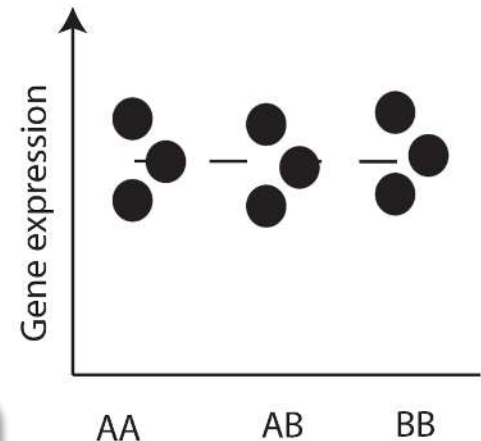
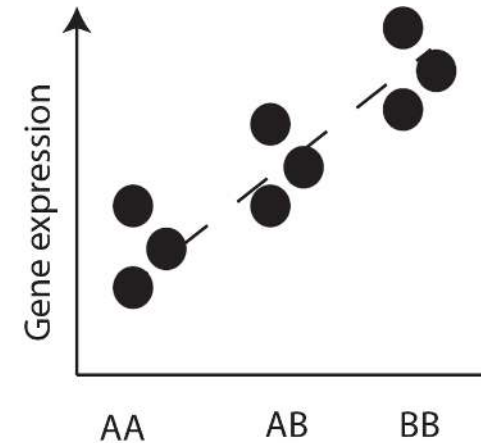
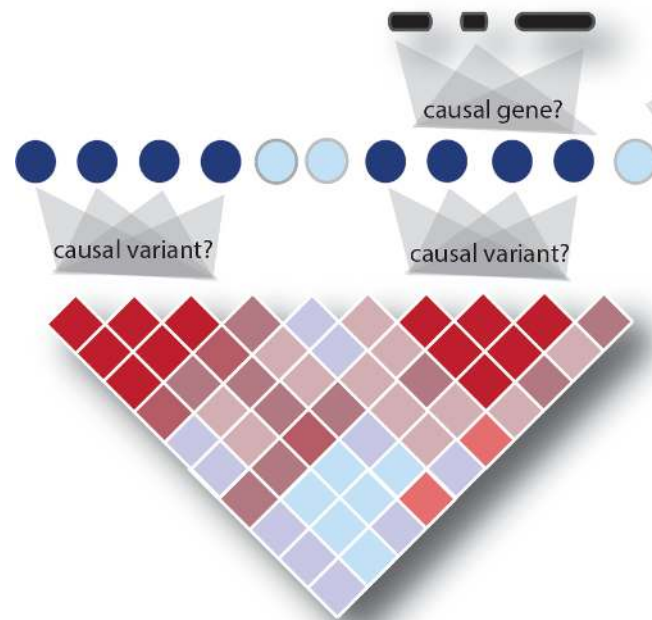
Associated variants fall in the regions of extended LD



The challenge to identify disease relevant genes

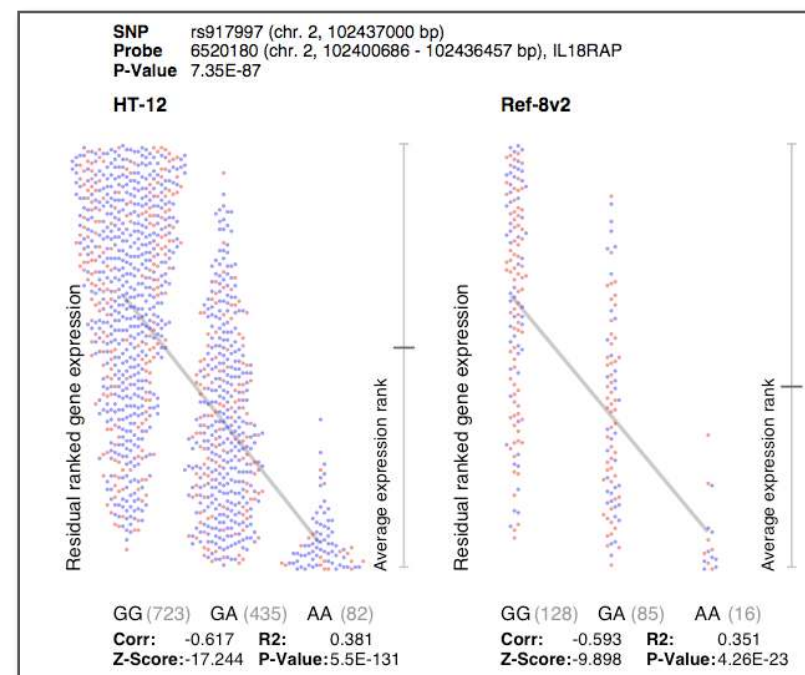
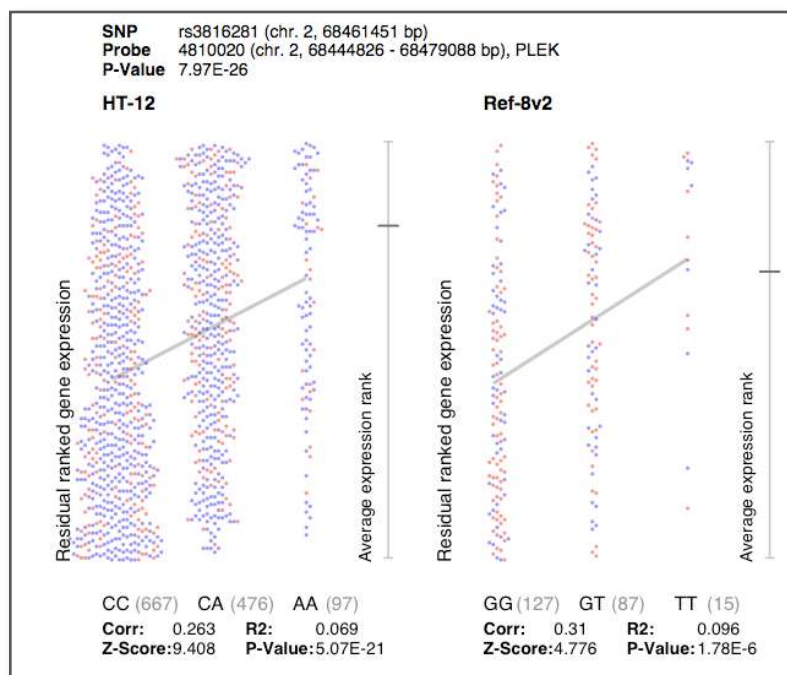


Correlating SNP genotypes with gene expression

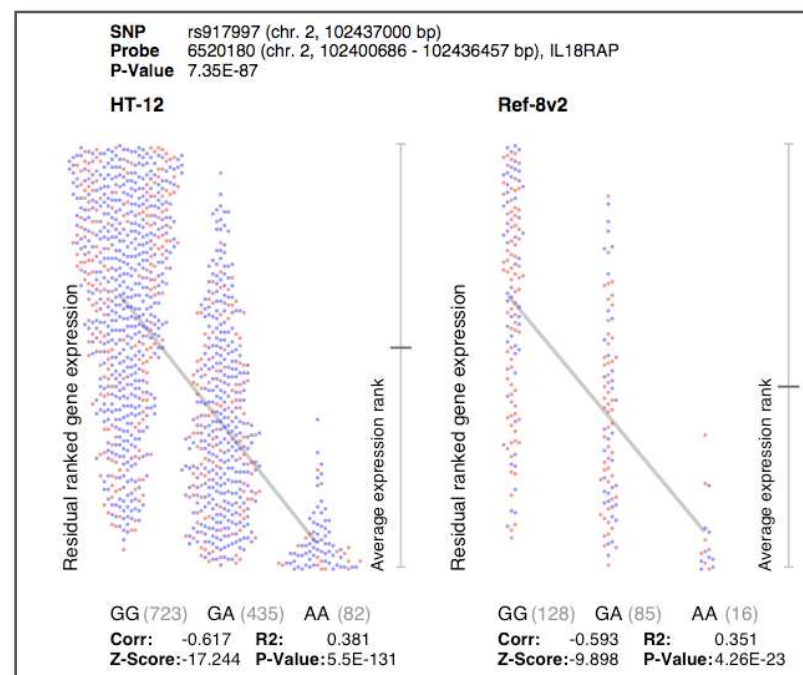
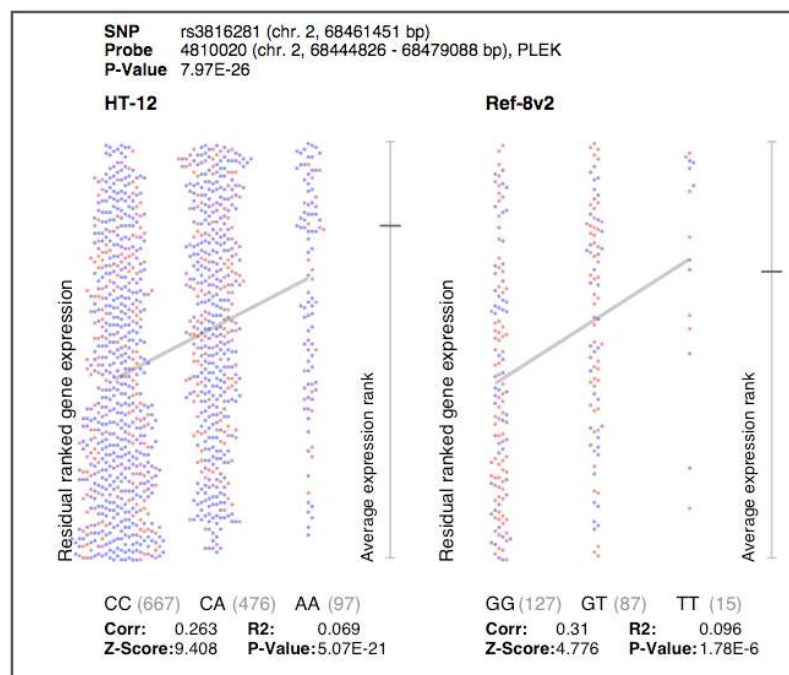


Expression quantitative
trait loci (eQTL)

Genotype-gene expression correlation across 1500 samples

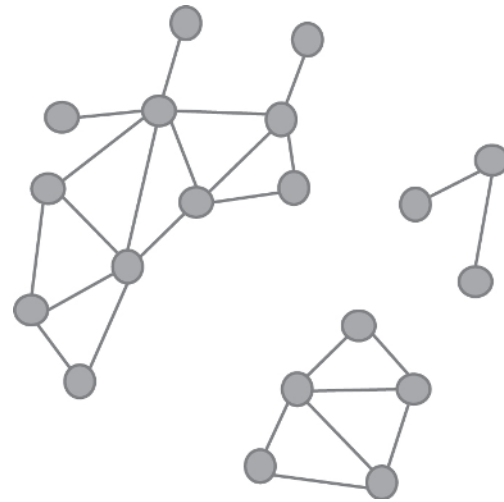


Genotype-gene expression correlation across 1500 samples

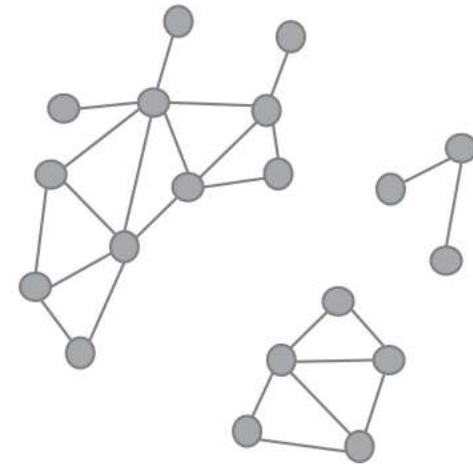
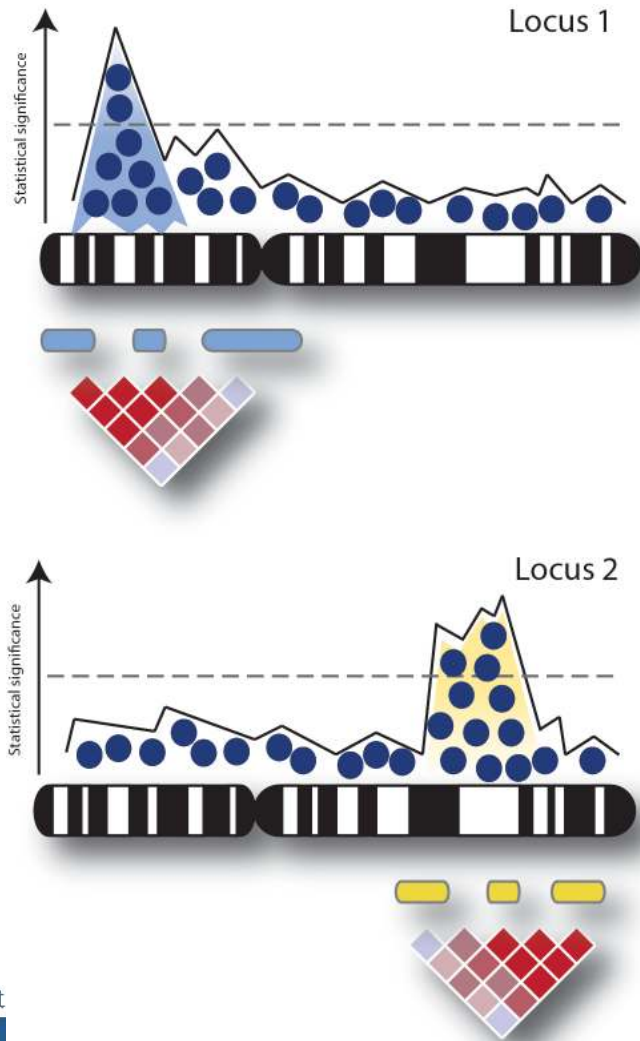


53% of associated variants influence gene expression

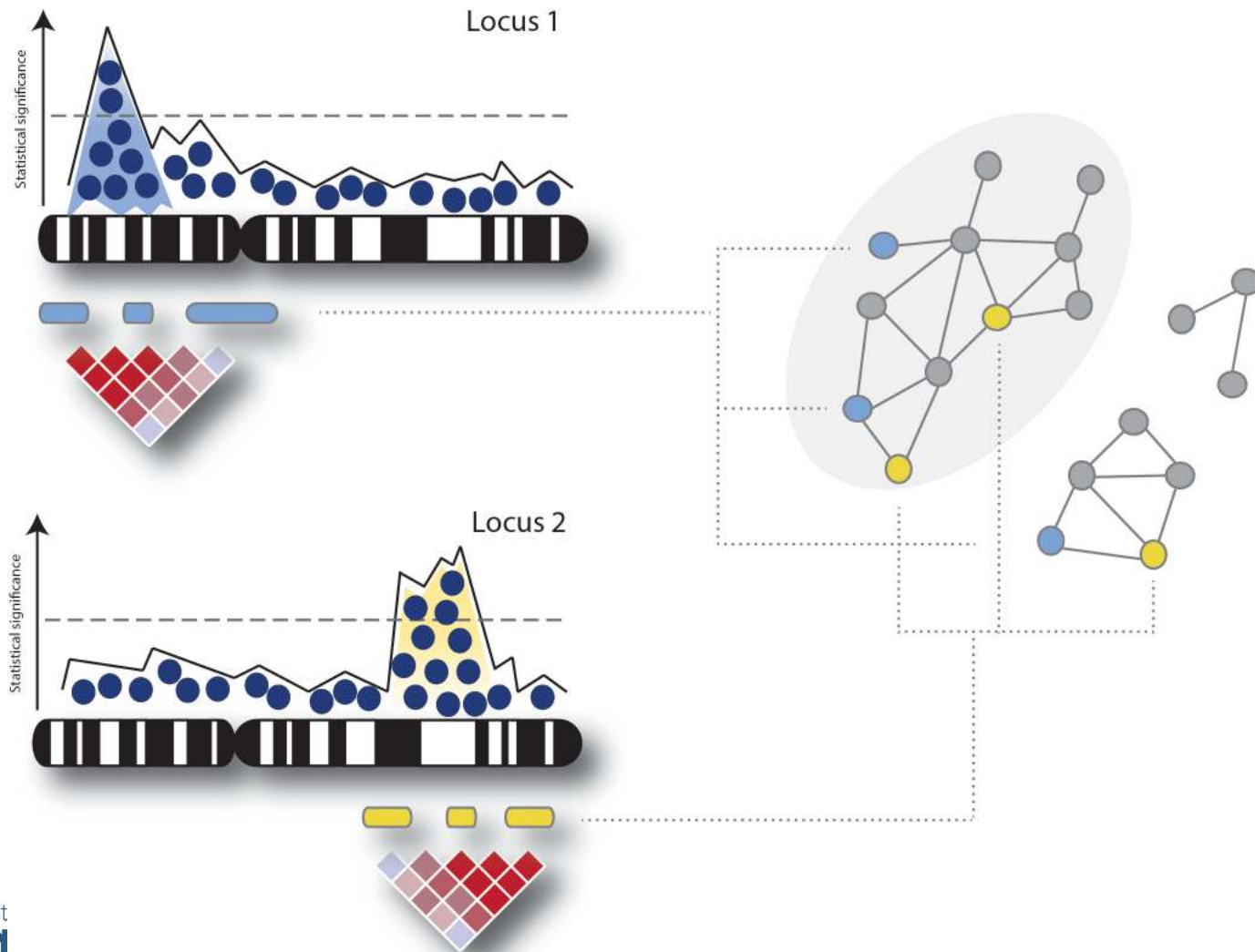
Genes co-expressed can indicate biologically functional modules

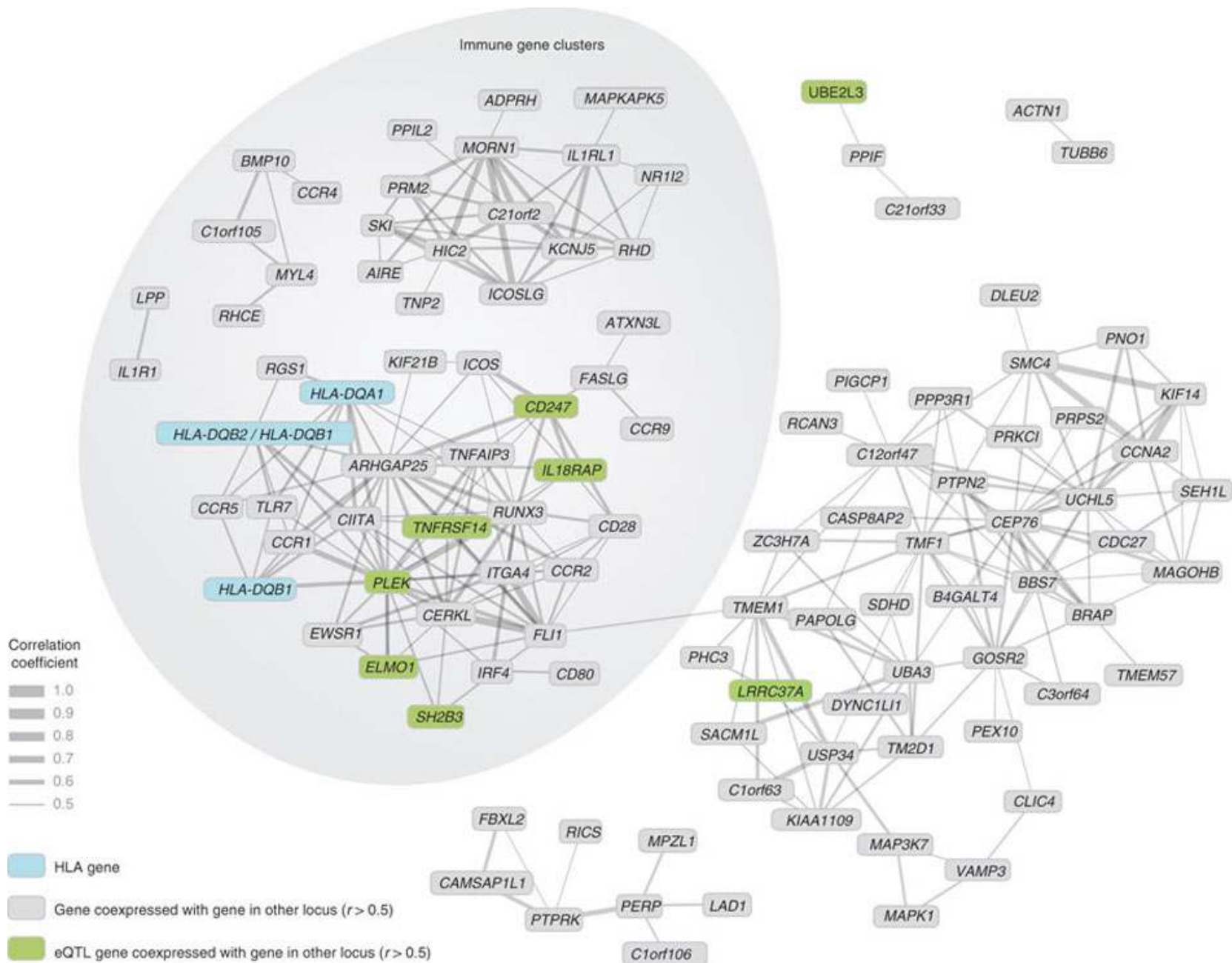


Genes co-expressed can indicate biologically functional modules



Genes co-expressed can indicate biologically functional modules





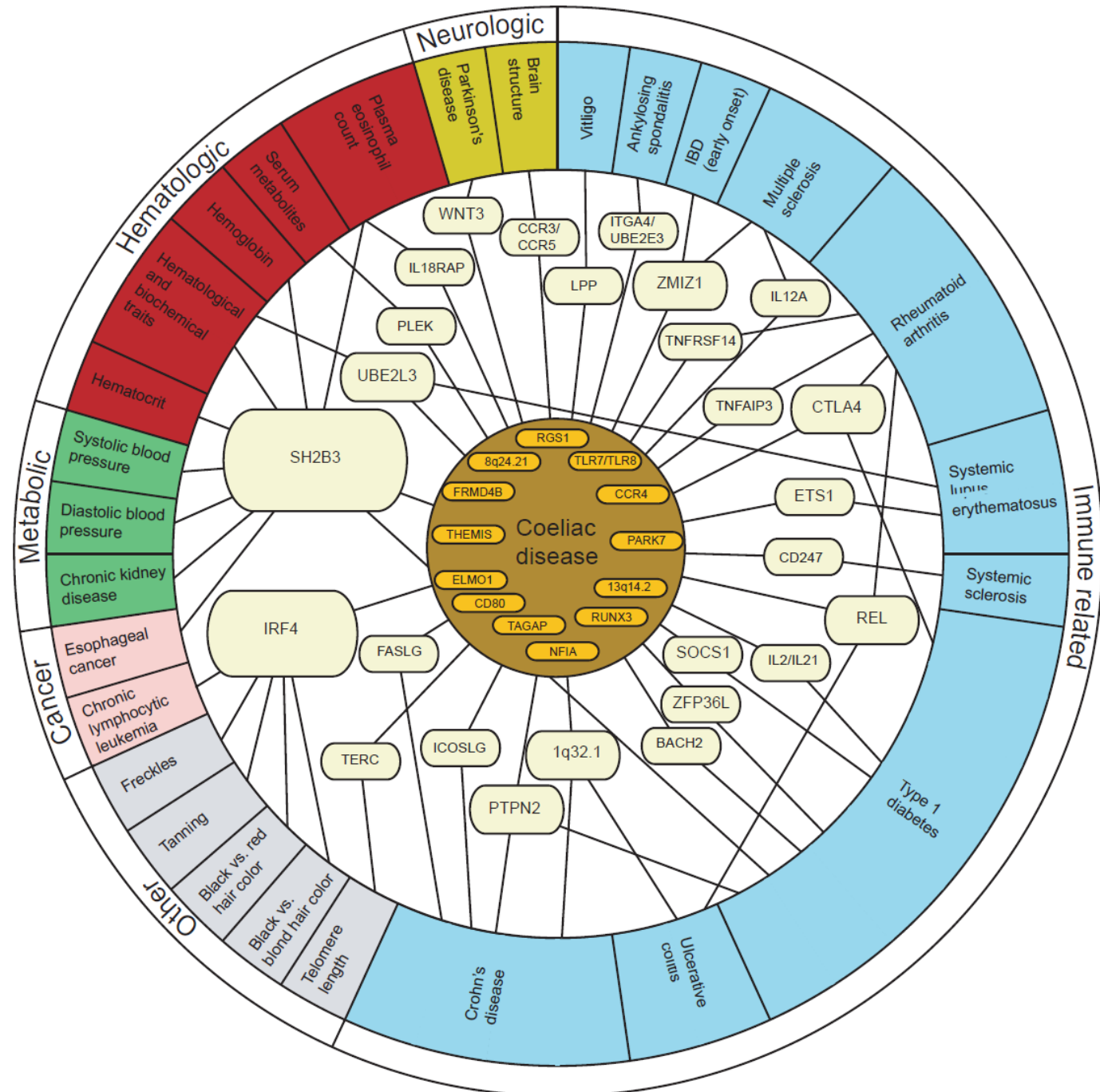
Lude Franke, UMCG, the Netherlands

<http://genenetwork.nl/>



Genetic sharing across immune diseases, motivation for ImmunoChip consortium

Associated loci are shared with other diseases



ImmunoChip

- Follow-up on results from 12 autoimmune GWASs
- 196,524 manufactured SNPs

Design

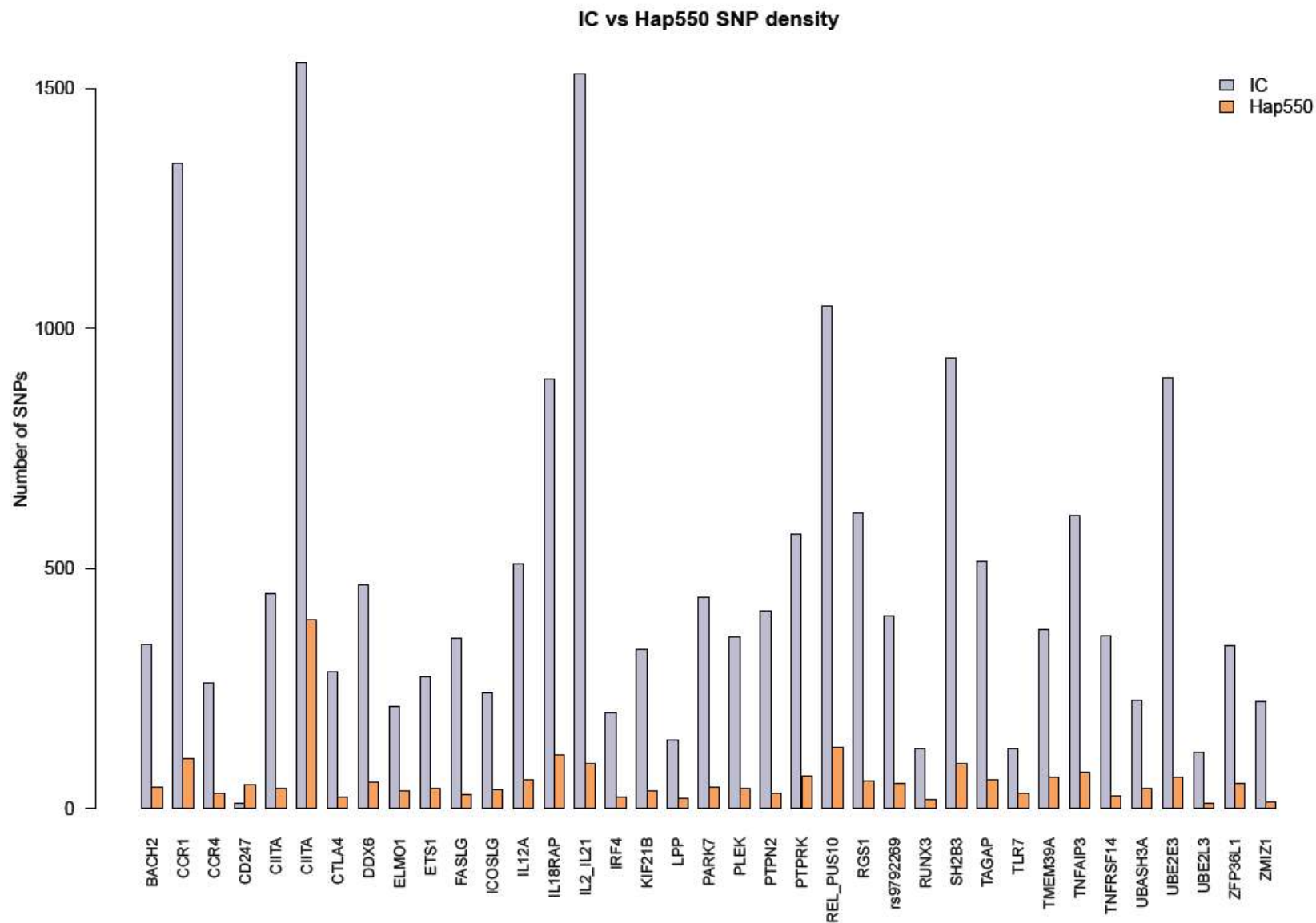
- Replication of signals with intermediate significance

Replication

- Dense genotyping at 186 loci associated to immune-related diseases (10-20x greater marker density)
- Sequence variants (early release of 1000 Genomes Project and individual sequencing efforts)

Fine-mapping

10-20x greater SNP density than Hap550 (post-QC)

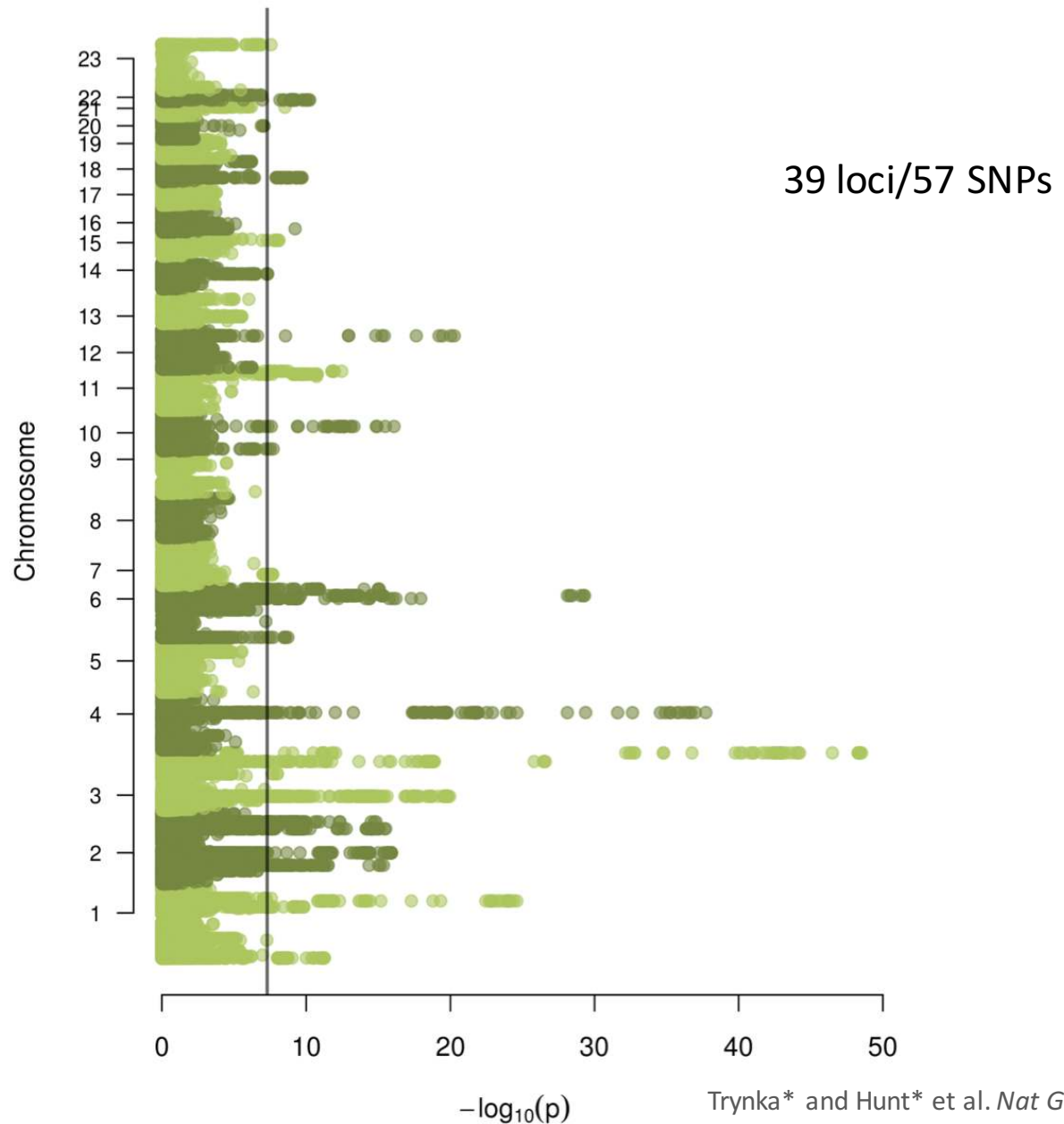


Sample collections – ultra clean dataset

Population	Celiac cases	Controls
UK	7728	8274 ^b
The Netherlands	1123	1147
Poland	505	533
Spain - Basque ^a	545	308
Spain - Madrid ^a	537	320
Italy - Rome, Milan, Naples	1374	1255
India - Punjab	229	391
Total	12041	12228

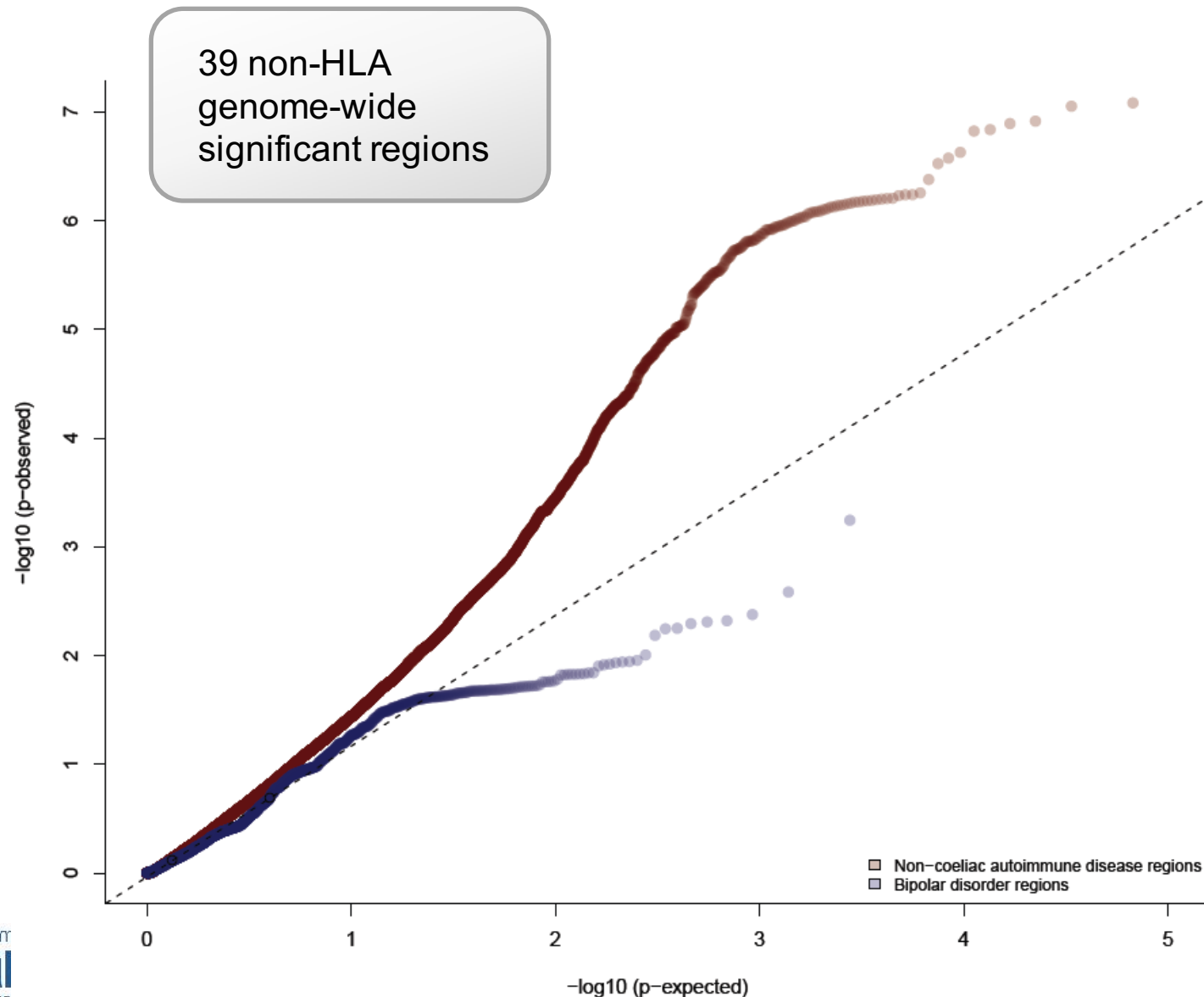
- Sample call rate > 99.5%
- SNP call rate > 99.95%
- Overall 0.008% missing genotype calls
- 139,553 polymorphic SNPs (excluded 15,657 NP variants)
- 24,661 variants: 5% > MAF > 0.05% and 22,941 MAF < 0.05%

Identification of 39 risk loci



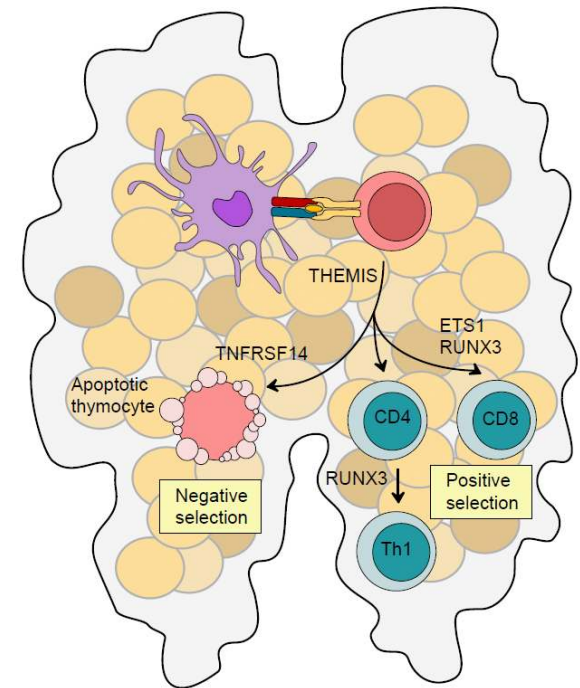
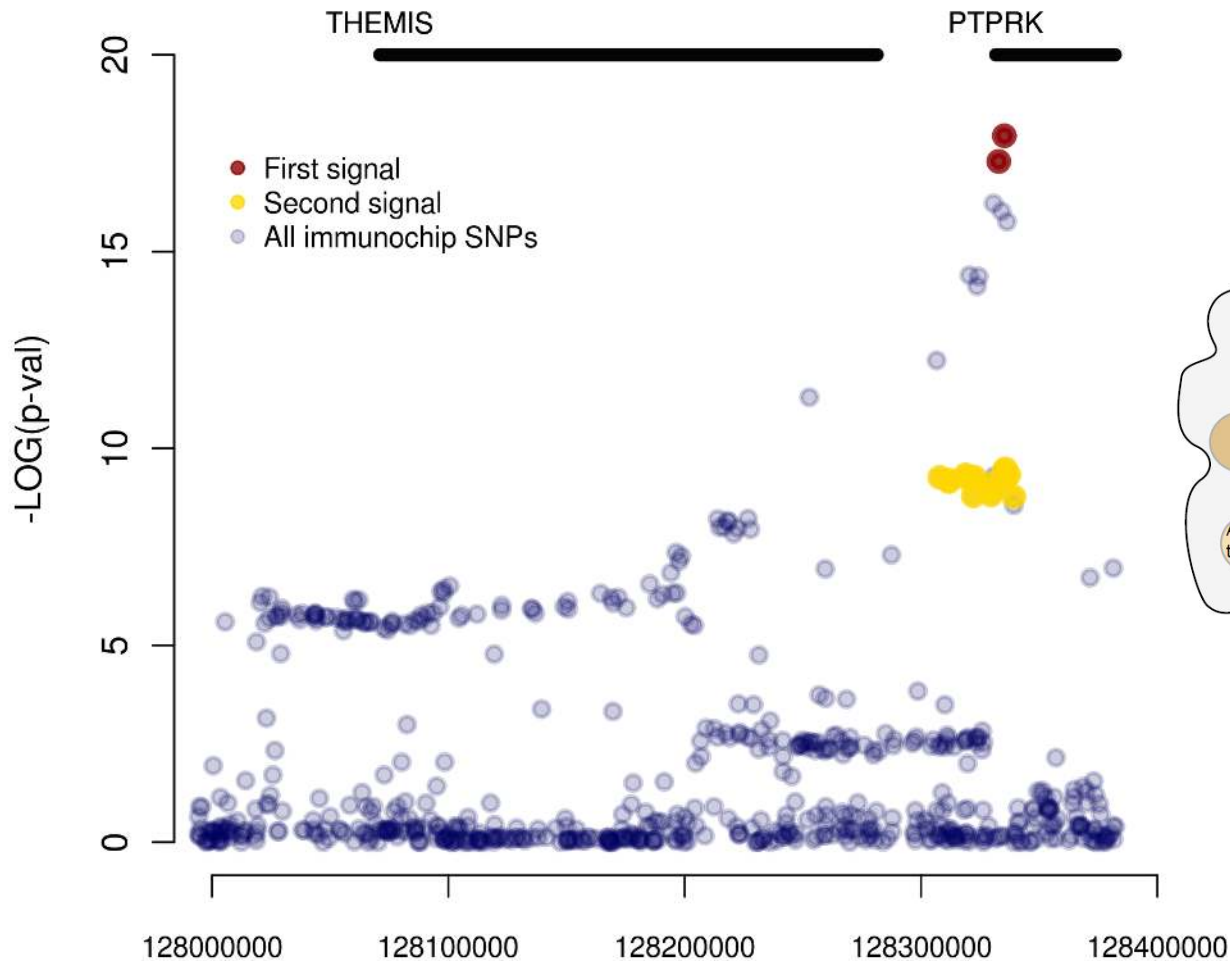
Trynka* and Hunt* et al. *Nat Genet*, 2011

Excessive enrichment for association signals in autoimmune loci

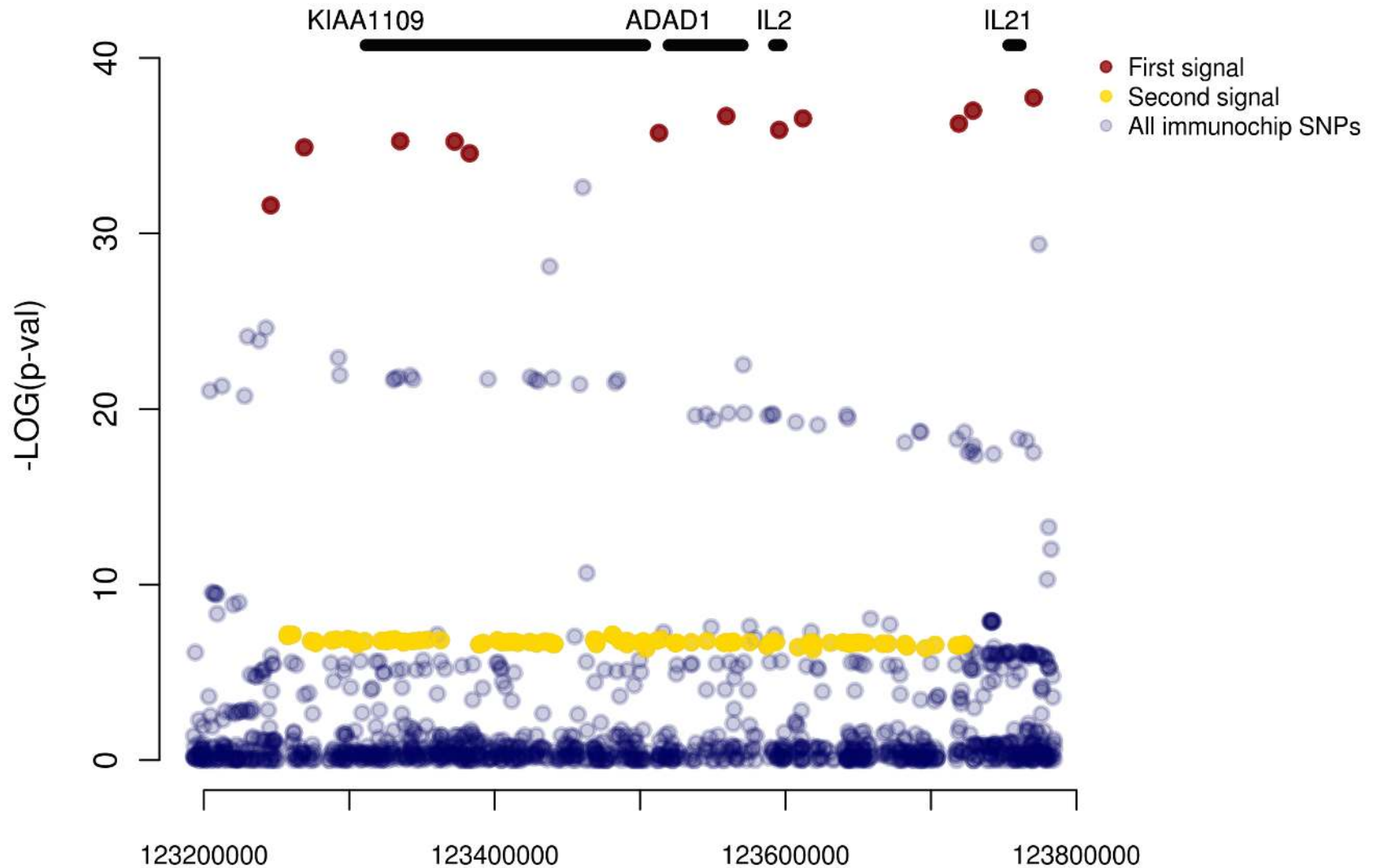


- Dense SNP map should allow us to carry out statistical fine-mapping to refine association signals

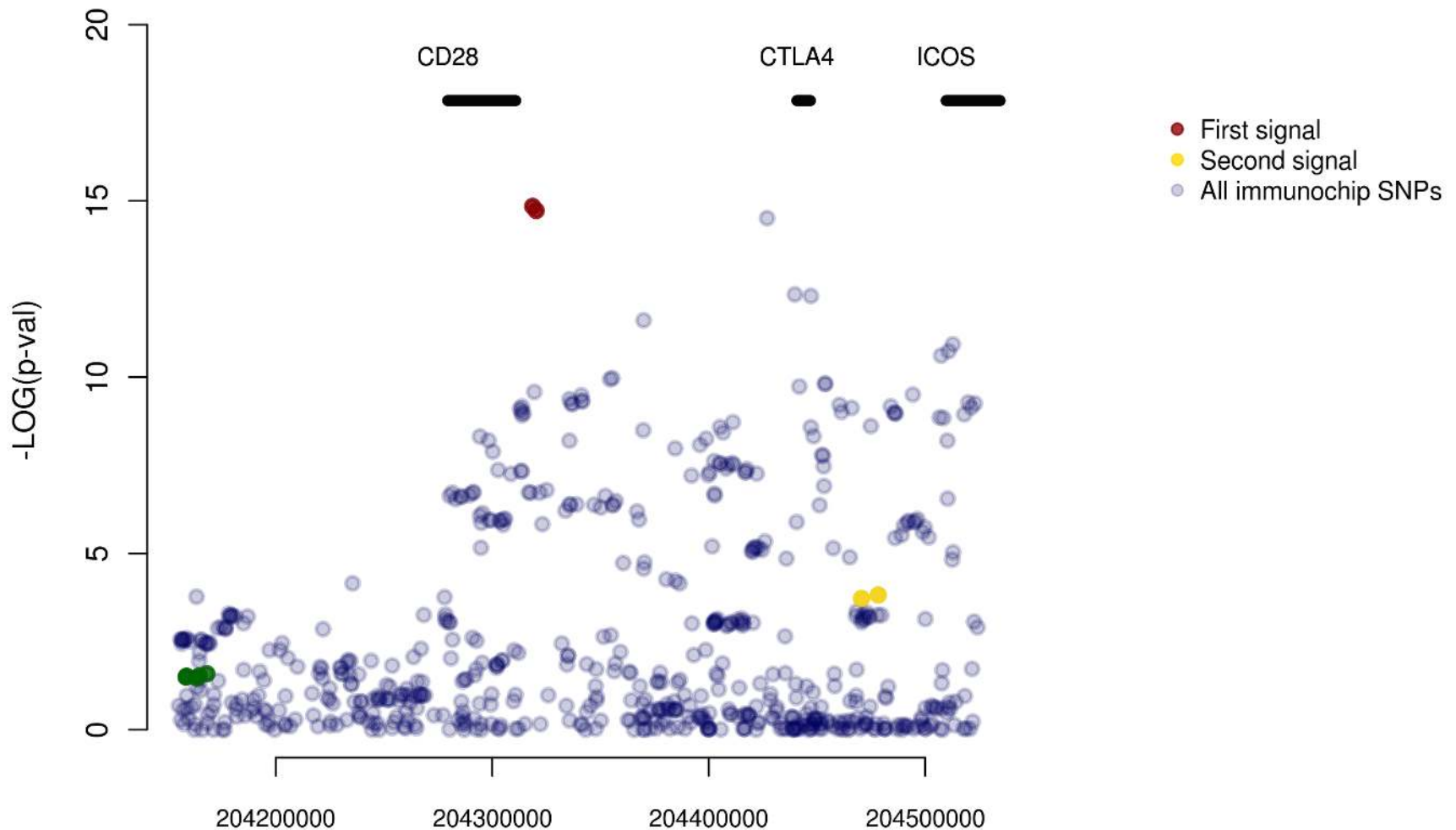
Refining the association signals

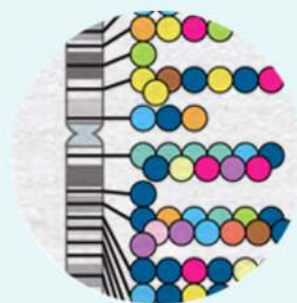


Unsuccessful fine-mapping



Multiple independent signals





GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies



Examples: breast cancer, rs7329174, Yang, 2q37.1, HBS1L

<https://www.ebi.ac.uk/fgpt/gwas/>



genome.gov

National Human Genome Research Institute

National Institutes of Health

[Research Funding](#)
[Research at NHGRI](#)
[Health](#)
[Education](#)
[Issues in Genetics](#)
[Newsroom](#)
[Careers & Training](#)
[About](#)
[For You](#)

[Home](#) > [Research Funding](#) > [Research Funding Divisions](#) > [Division of Genomic Medicine](#) > [GWAS Catalog](#)

[Share](#) [Print](#)

Division of Genomic Medicine

A Catalog of Published Genome-Wide Association Studies

[Division Staff](#) : [Funding Opportunities](#) : [Genomic Medicine Activities](#) : [GWAS Catalog](#) : [Meetings & Workshops](#) : [Potential Sample Collections for Sequencing](#) : [Programs](#) : [Publications](#) : [Trans-NIH Sequencing Inventory](#)

Additional information has been added to the HTML catalog columns below. For a description of column headings for the HTML catalog, go to: [Catalog Heading Descriptions](#)

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

[View the Full Catalog](#)
[Download the Catalog](#)
[Search the Catalog](#)

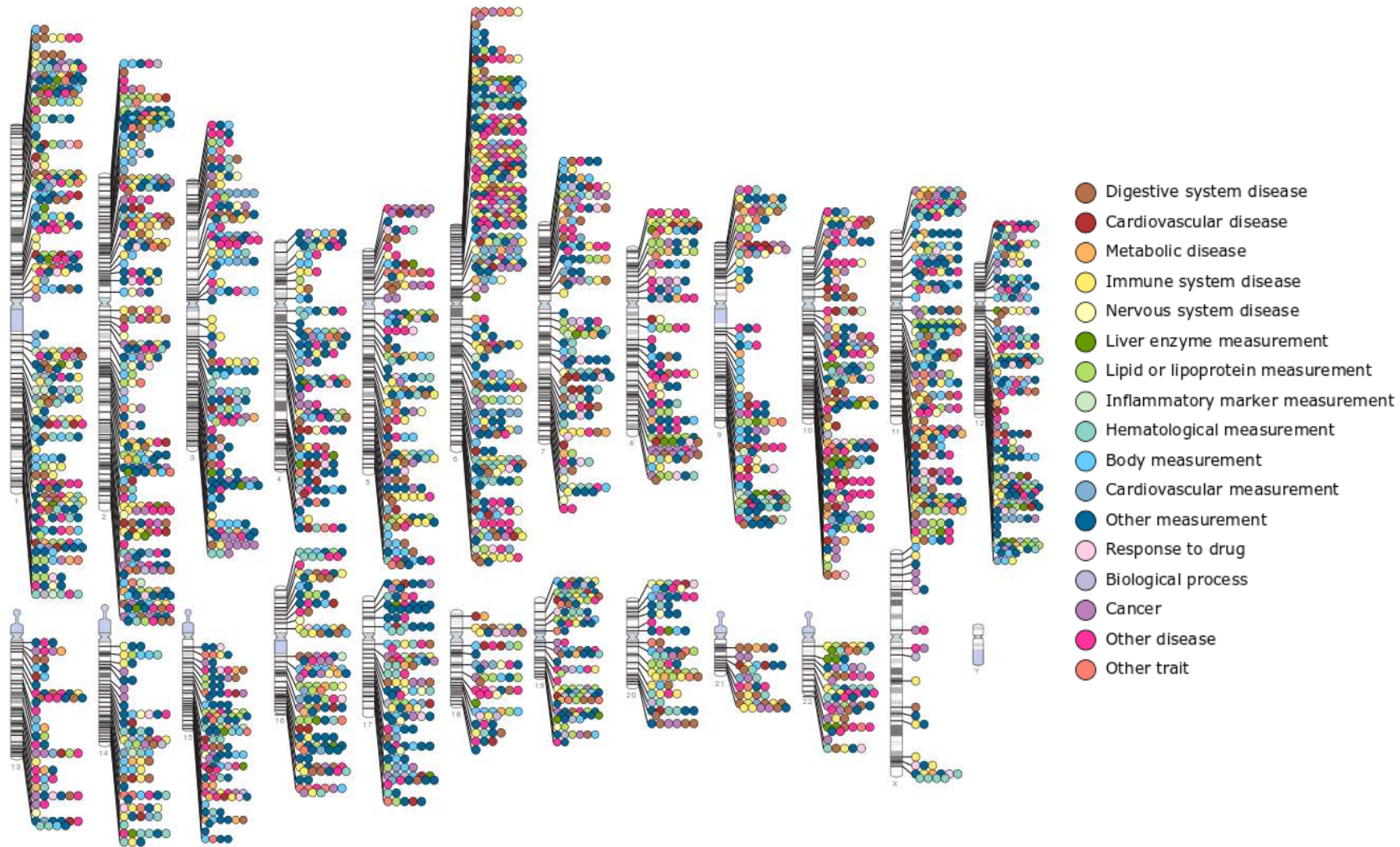


The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS literature ([HuGE Navigator](#)).

SNP-trait associations listed here are limited to those with p-values < 1.0×10^{-5} (see full methods for additional details). Multipliers of powers of 10 in p-values are rounded to the nearest single digit; odds ratios and allele frequencies are rounded to two decimals. Standard errors are converted to 95 percent confidence intervals where applicable. Allele frequencies, p-values, and odds ratios derived from the largest sample size, typically a combined analysis (initial plus replication studies), are recorded below if reported; otherwise statistics from the initial study sample are recorded. For quantitative traits, information on % variance explained, SD increment, or unit difference is reported where available. Odds ratios < 1 in the original paper are converted to OR > 1 for the alternate allele. Where results from multiple genetic models are available, we prioritized effect sizes (OR's or beta-coefficients) as follows: 1) genotypic model, per-allele estimate; 2) genotypic model, heterozygote estimate, 3) allelic model, allelic estimate.

<https://www.genome.gov/gwastudies/>

We have mapped thousands of disease loci!

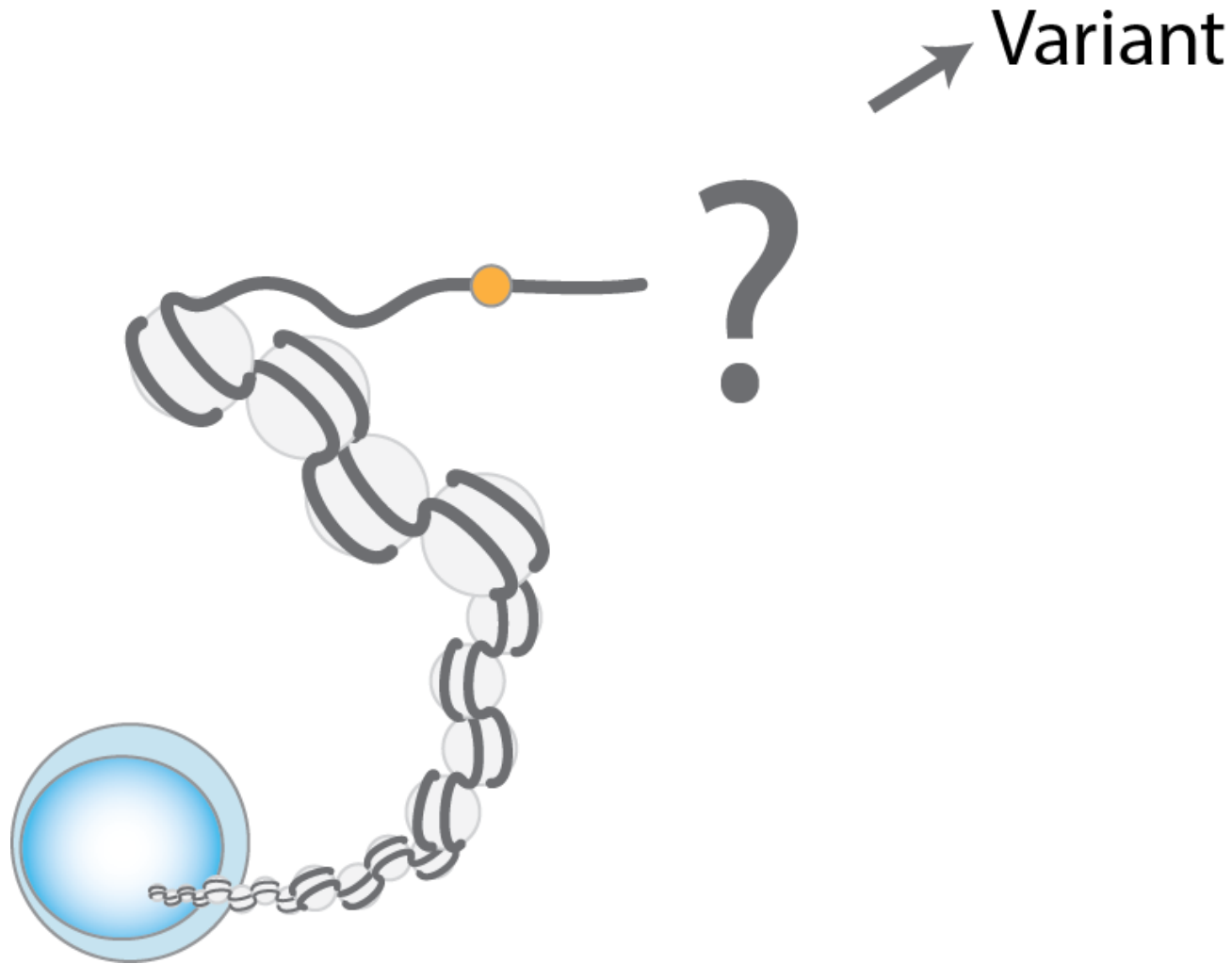


And then we got stuck...

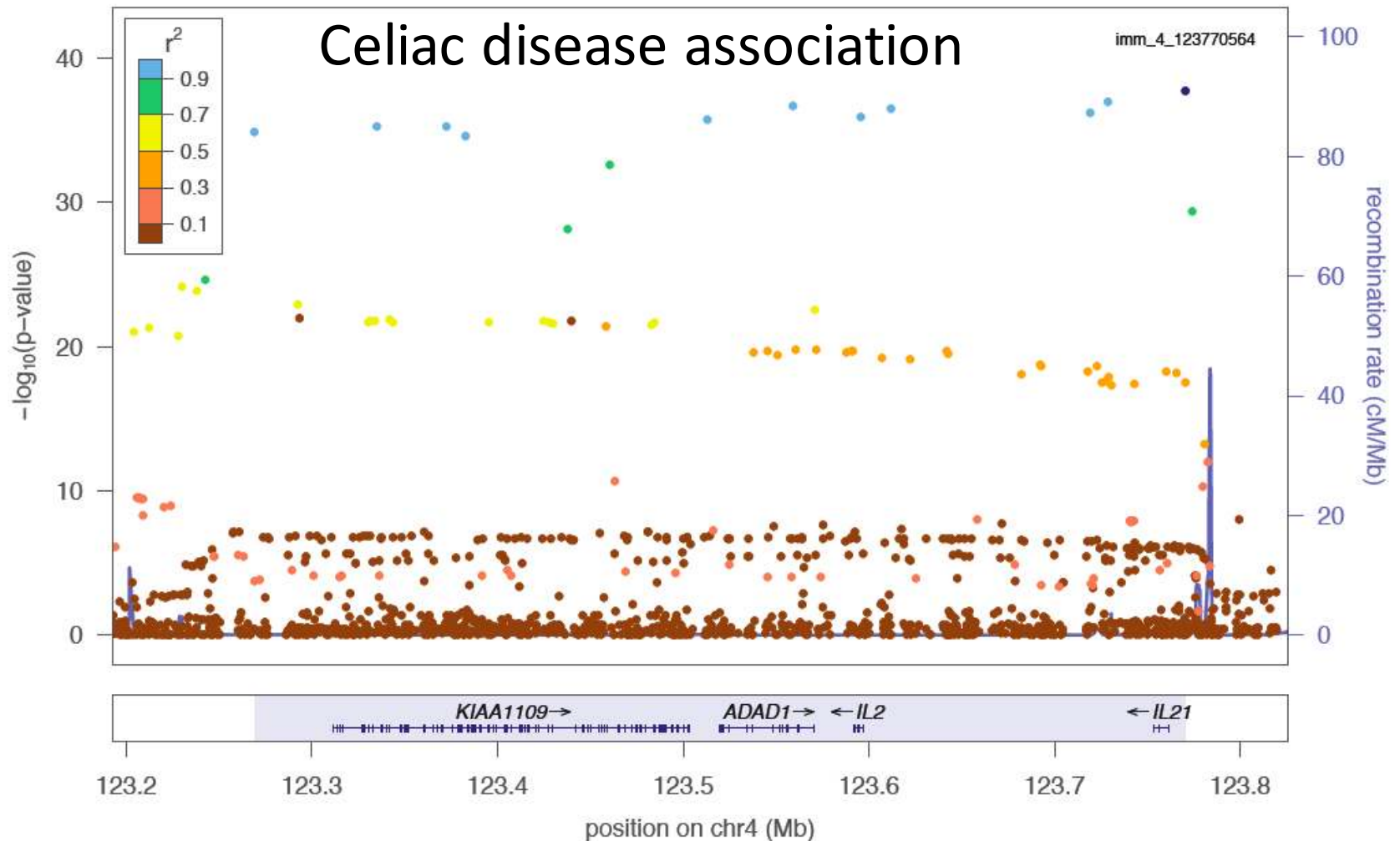
The path from
disease association
to function remains
a challenge



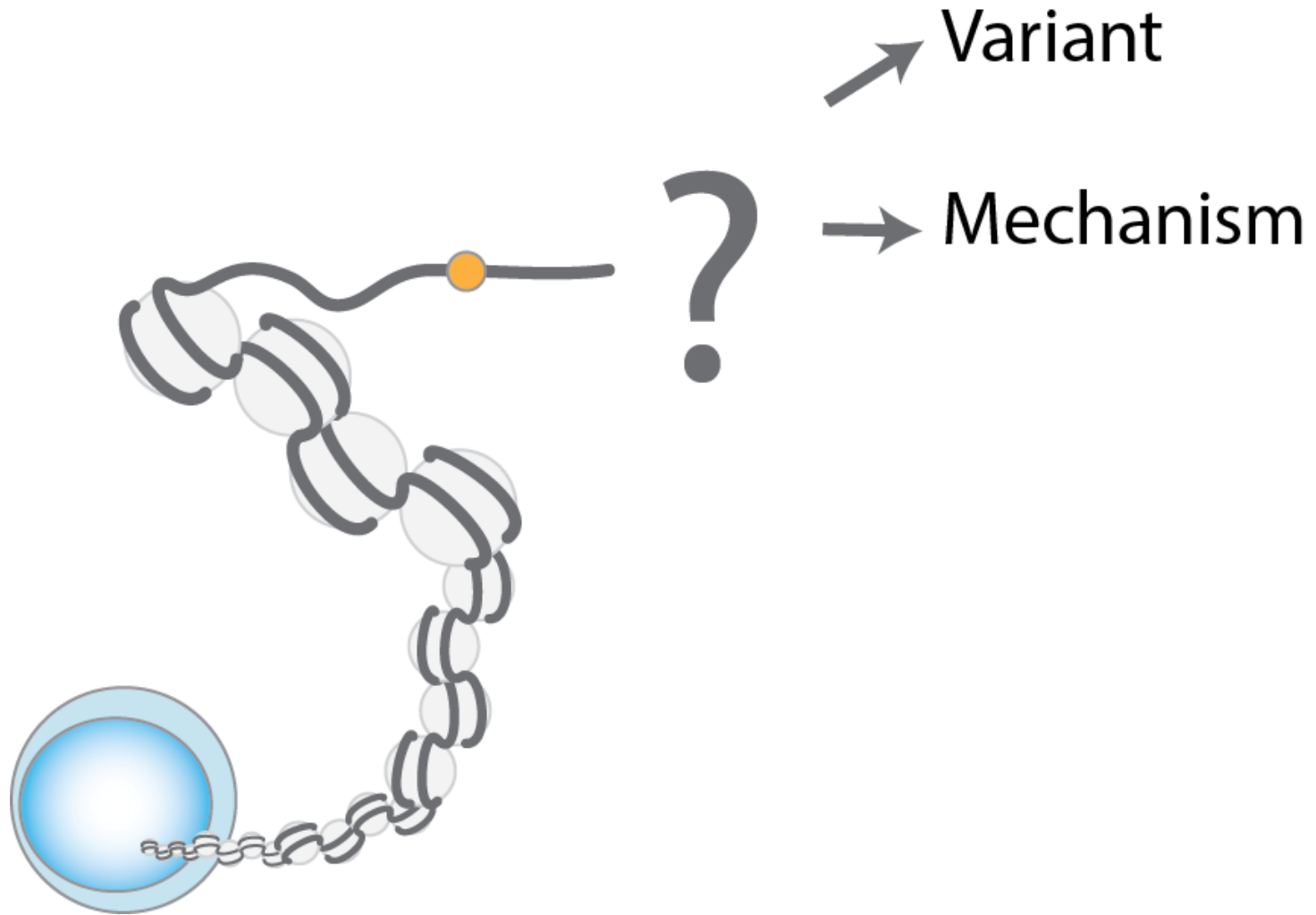
Interpreting disease variants



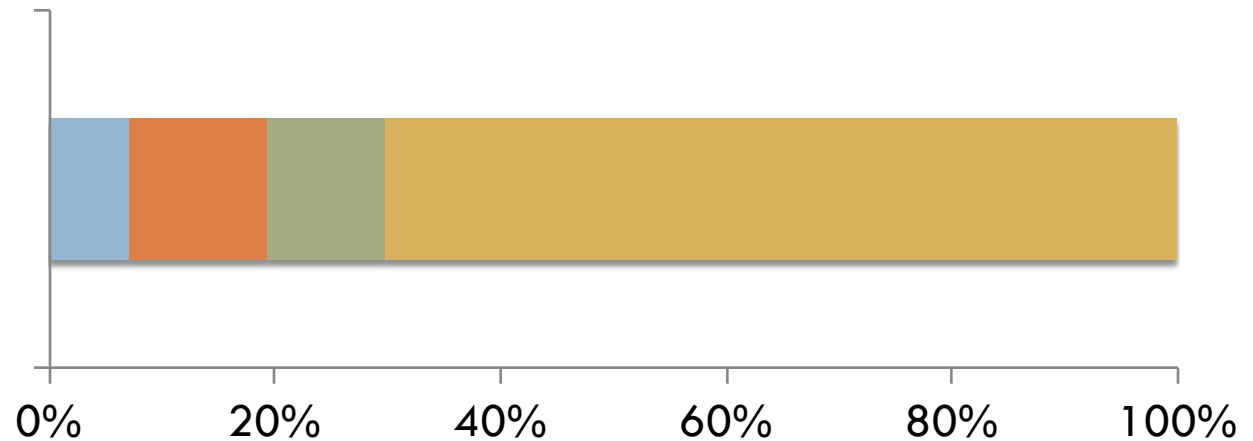
The challenge to identify relevant variants – extended LD



Interpreting disease variants



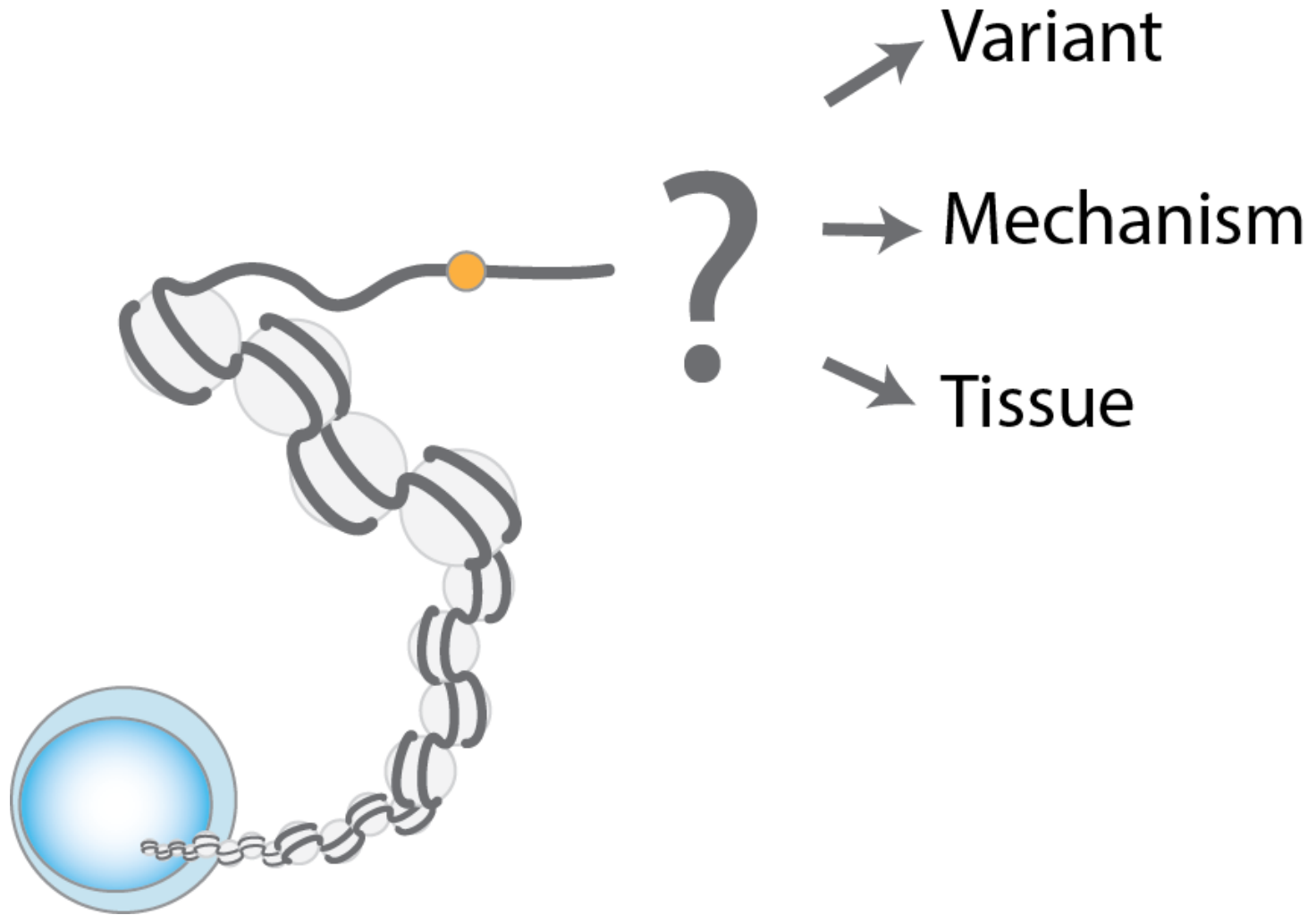
Celiac disease alleles implicate gene regulatory function



39 loci/57 SNPs

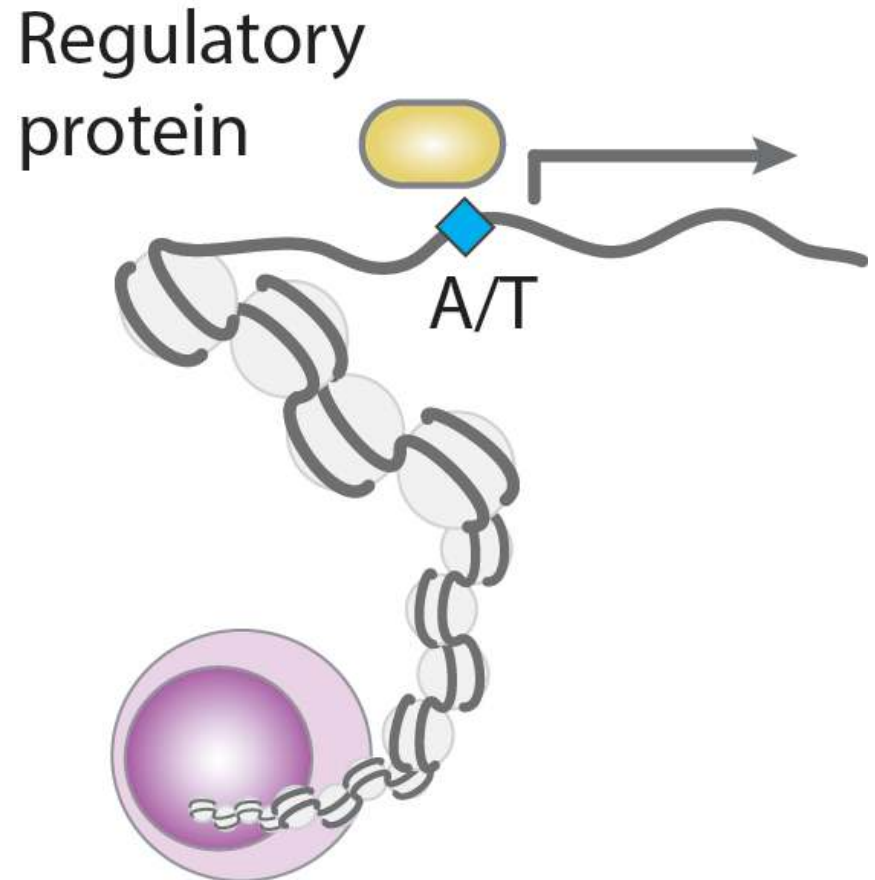
- Coding
- Transcribed Non-Coding
- Regulatory
- Other

Interpreting disease variants

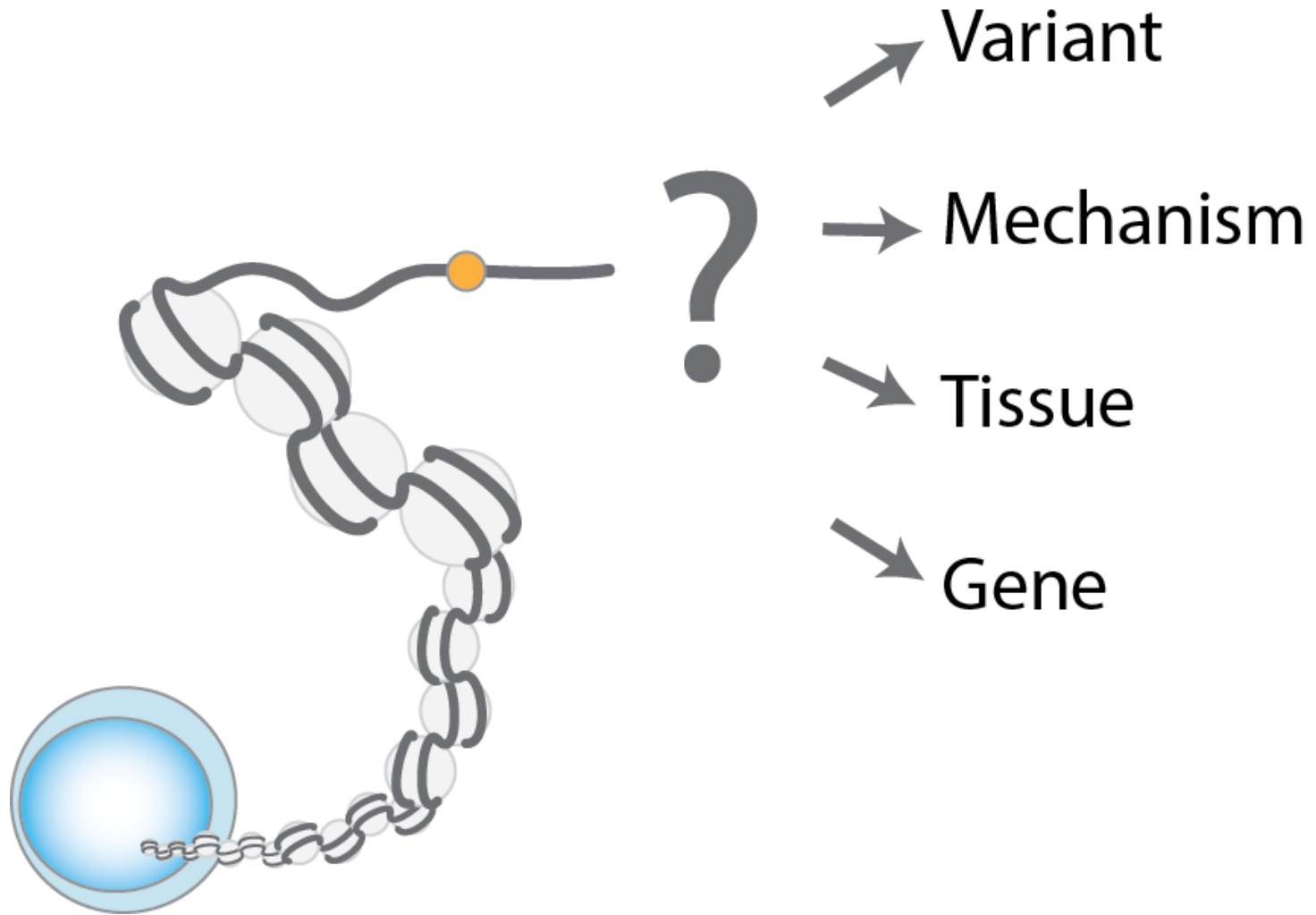


A handful of studies describing function of non-coding disease variants

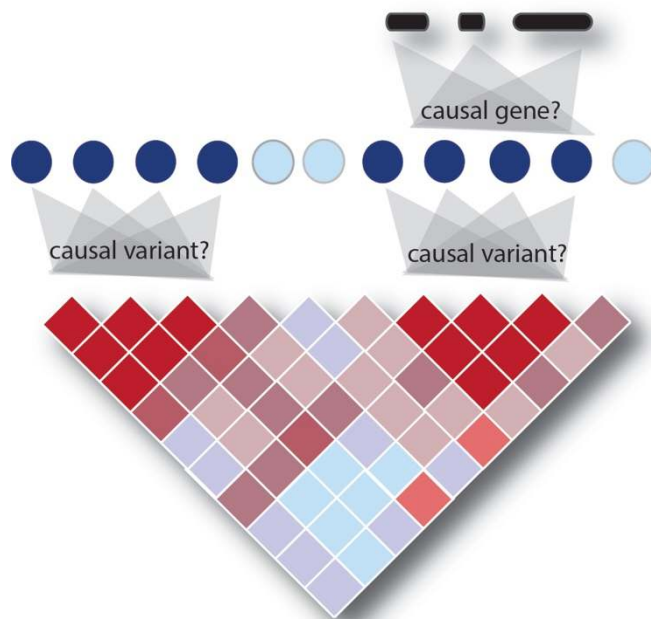
- E.g. LDL (Musunuru et al., *Nature* 2010)
- LDL associated non-coding variant creates transcription factor binding site and alters the hepatic expression of the *SORT1*



Interpreting disease variants

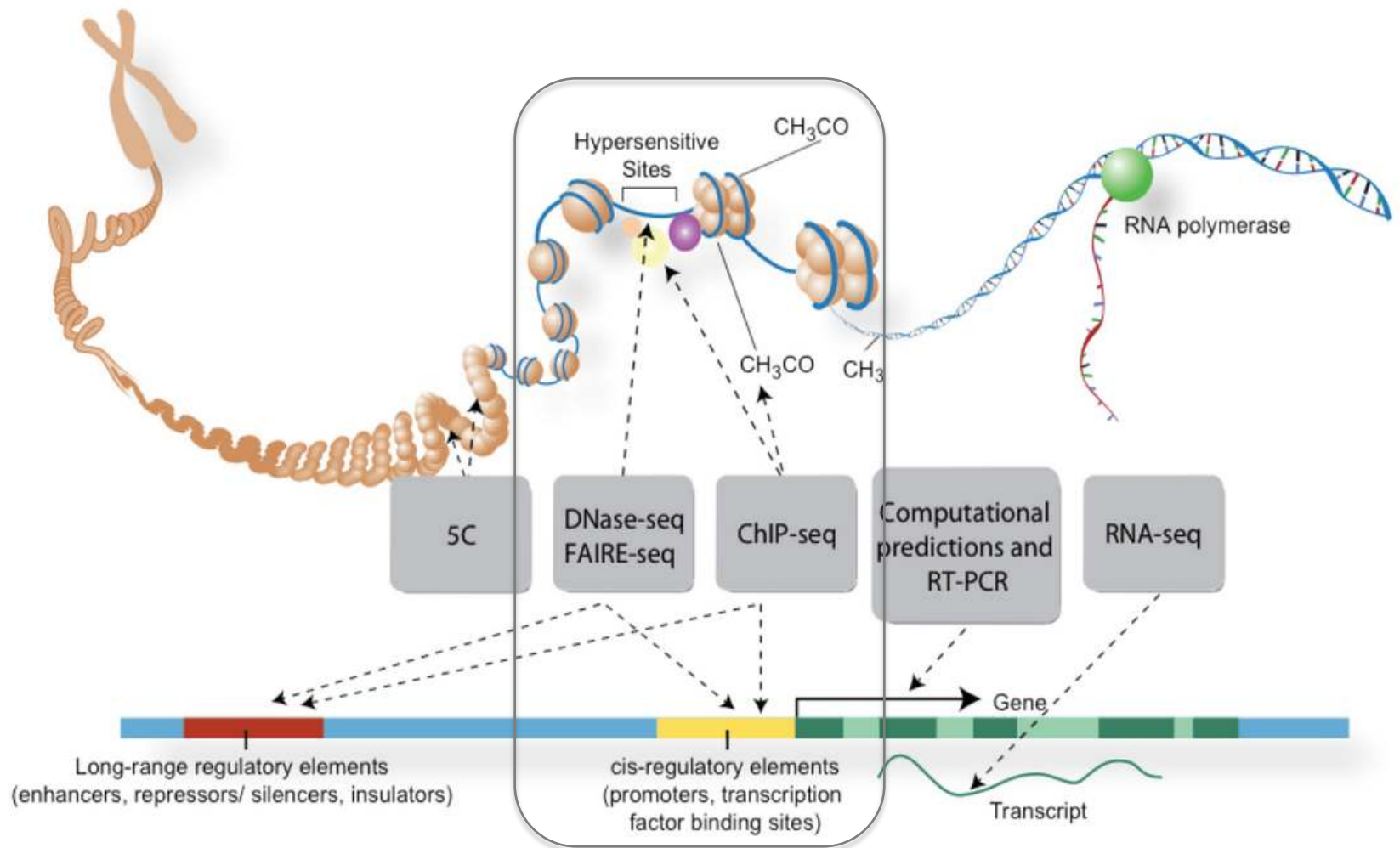


The challenge to identify relevant variants and genes

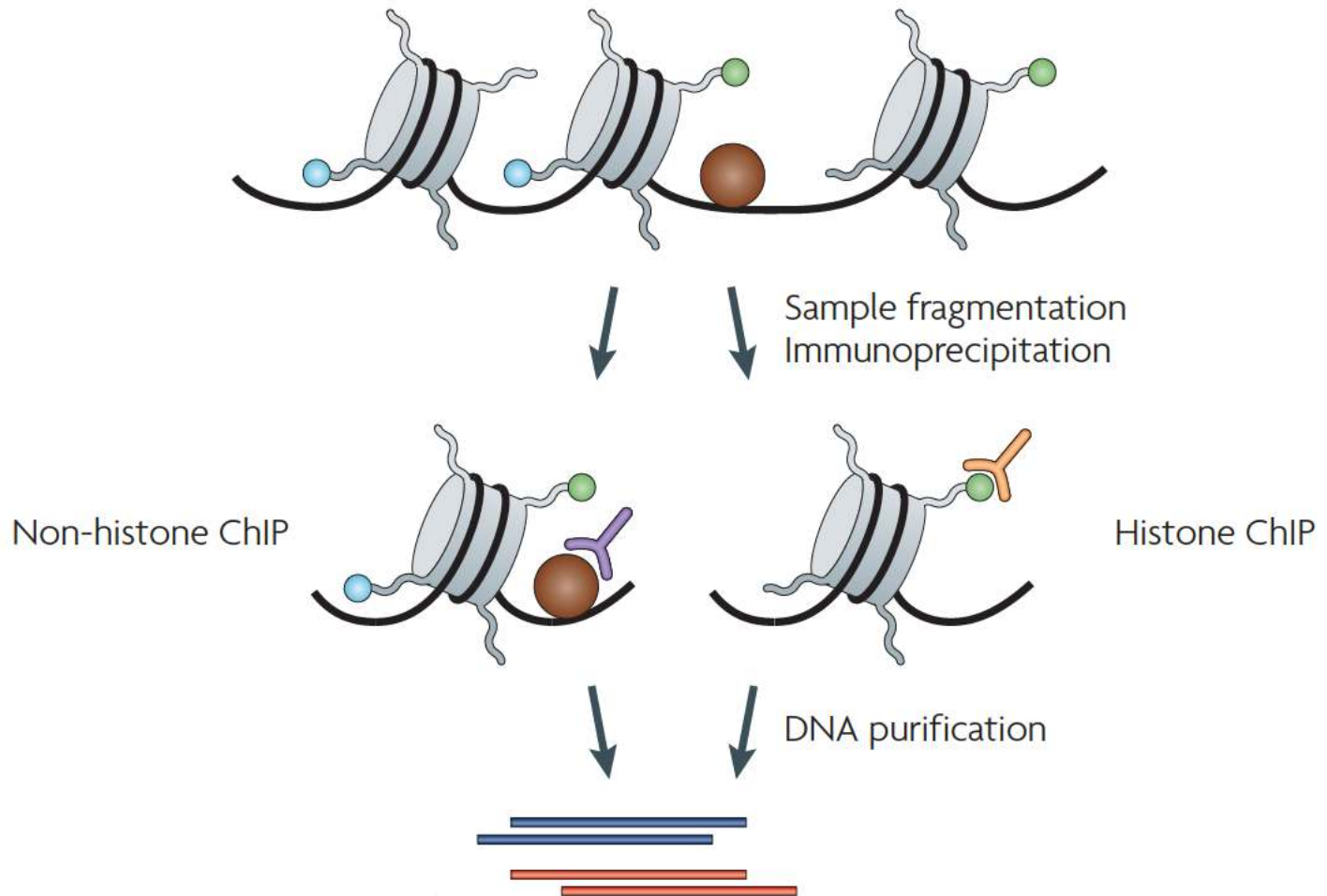


- Multiple genes within associated LD block

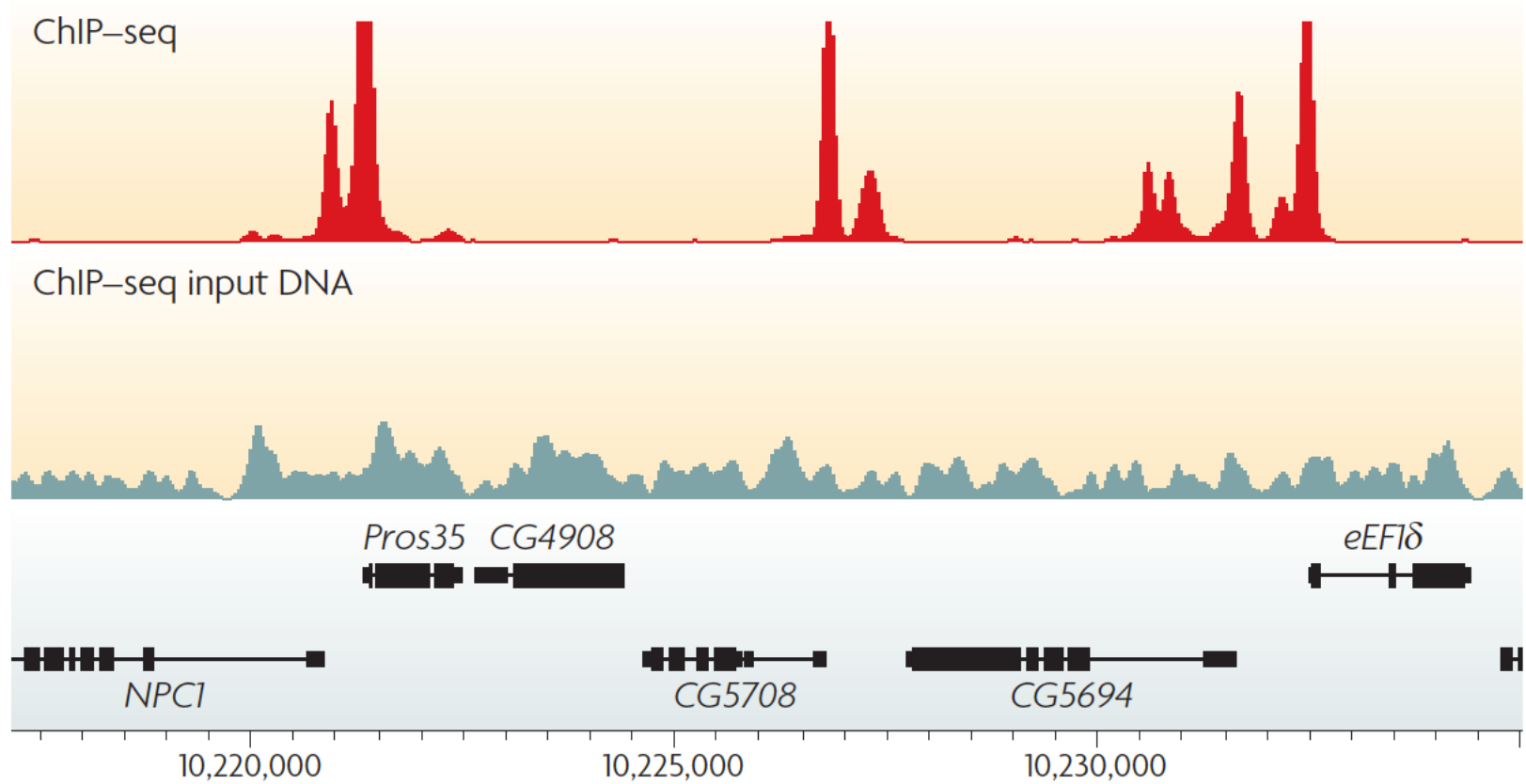
Mapping gene regulation



ChIP-seq



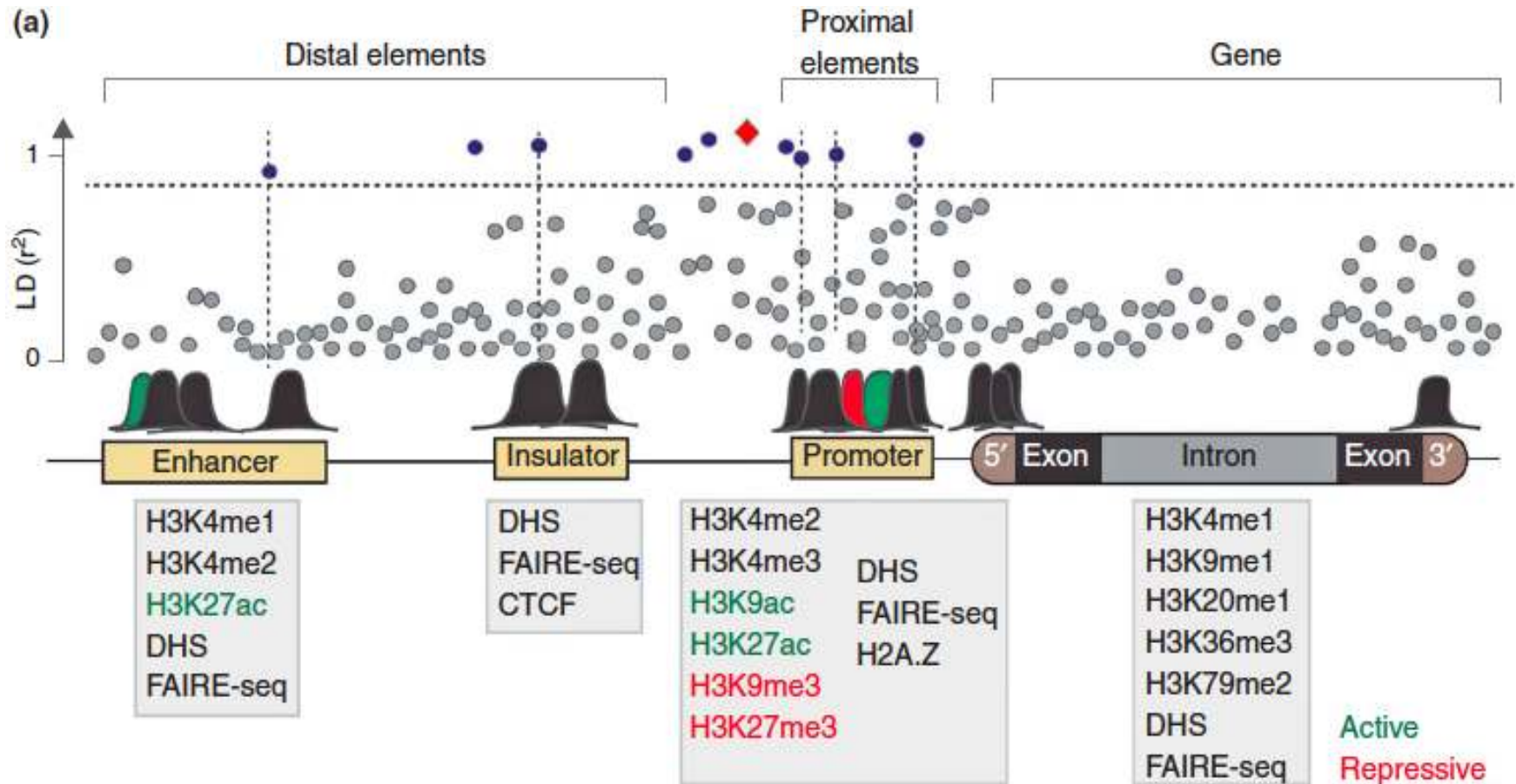
ChIP-seq



Resources

- ENCODE: cell types and cell lines, broad range of assays
- NIH Roadmap Epigenome Project: human tissue atlas, limited number of assays
- BLUEPRINT epigenome: >50 blood cells from healthy and patient samples, variome

Can we interestect associated variants with epigenetic data to learn about their function?



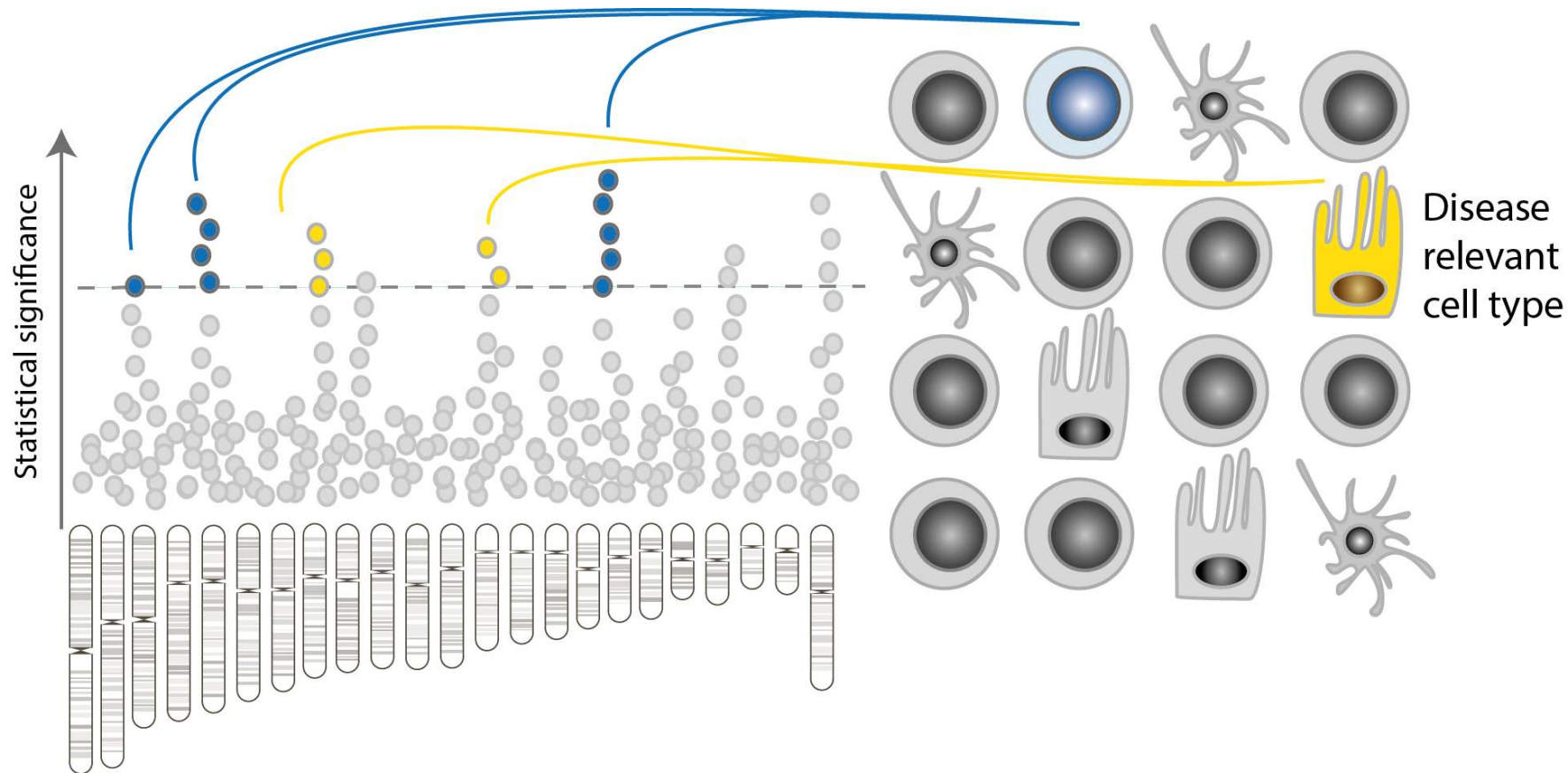
Challenges to functionally follow-up GWAS variants

- For many diseases the causal cell type is unknown
- Disease variants can express functionality in a cell type and a cell state specific context

Outline

- Identifying informative chromatin marks to perform quantitative genomic assays at scale
 - Trynka et al., Nature Genet 2013
- Developing a robust test to infer enrichment of associated variants across a range of genomic annotations
 - Trynka et al., AJHG 2015
- Where pathways and histone marks meet: increasing specificity to identify disease relevant cell types

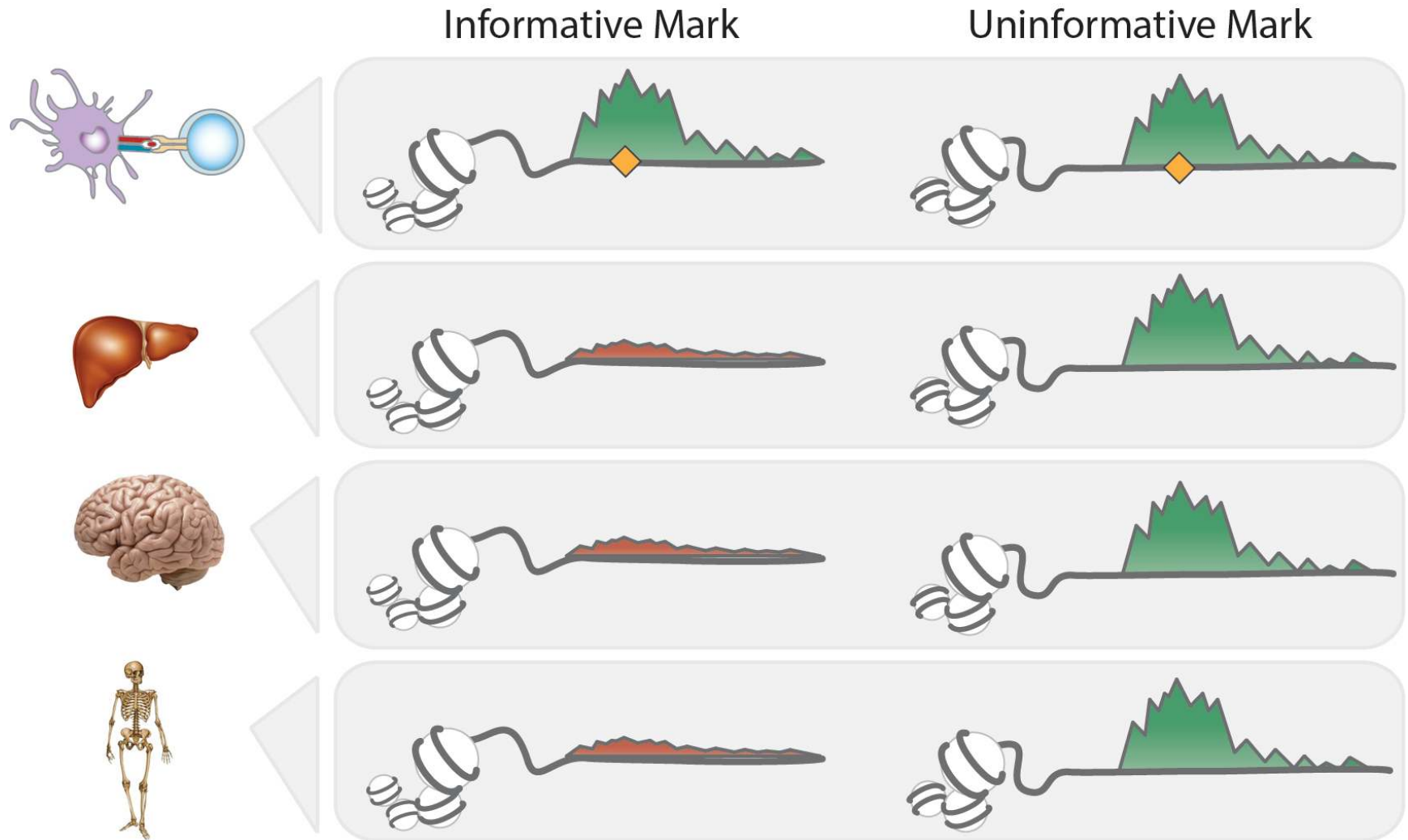
Leveraging cell type specific gene regulation with GWAS SNPs can pinpoint relevant cell types



Motivation

- 1) Is there a single informative chromatin mark?
- 2) Can it be used to:
 - identify cell-types relevant to the phenotype
 - fine-map to suggest causative variant

Cell-type specific signature of disease associated variants



Datasets – publically available

ENCODE:

- n = 14 cell lines
- 15 histone marks

The screenshot shows the ENCODE website with the title "Encyclopedia of DNA Elements". It includes a navigation menu on the left with links like "General", "Resources & FAQ", "Publications", "Software Tools", "Data Standards", "Human", "Downloads", "Experiment Matrix", "Search", and "Genome". The main content area has a section titled "About ENCODE Data" which describes the consortium's goal to build a comprehensive list of functional elements in the human genome. It also lists ways to use the data: "Search for displayable tracks and downloadable files", "Download of data files", "Visualization in the UCSC Genome Browser", and "Data mining with the UCSC Table Browser".

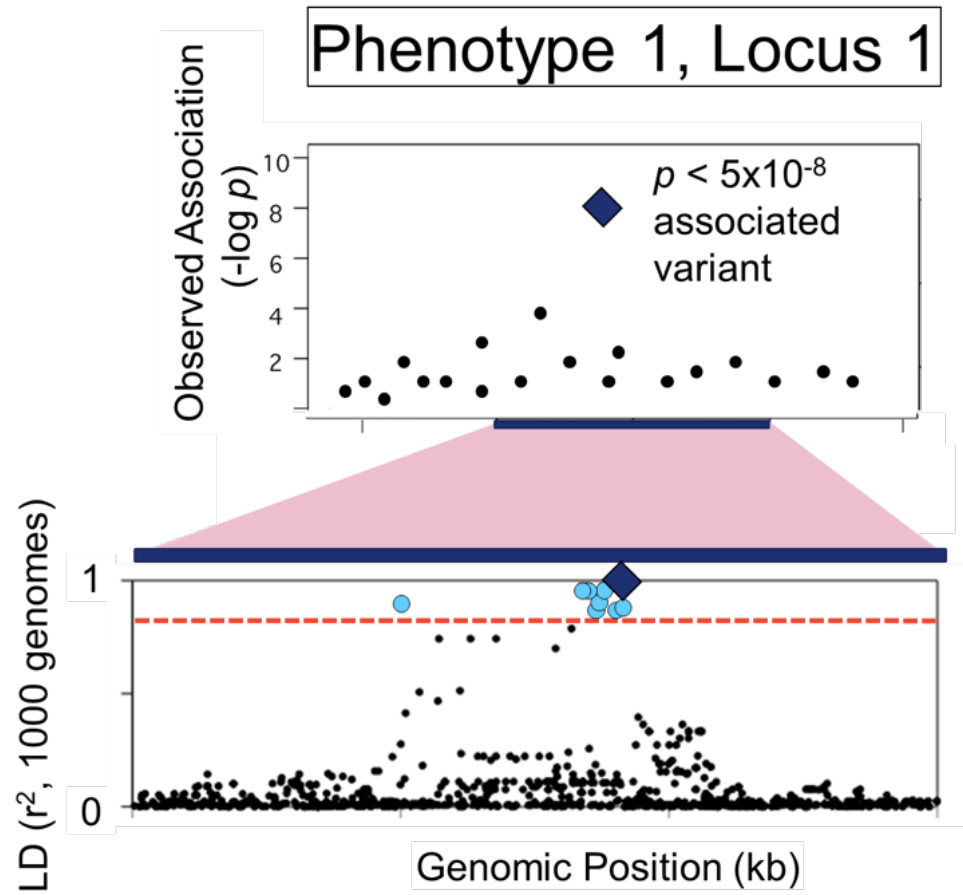
NIH Epigenome Project:

- n = 34 primary cells
- 6 histone marks

The screenshot shows the NIH Roadmap Epigenomics Project website. It features a navigation bar with links like "HOME", "PARTICIPANTS", "BROWSE DATA", "PROTOCOLS", "COMPLETE EPIGENOMES", "TOOLS", and "PUBLICATIONS". Below this is a section with tabs for "OVERVIEW", "PROJECT DATA", "MAPPING CENTERS", "PROTOCOLS & STANDARDS", "PUBLICATIONS", and "NEWS". The main content area includes a diagram of a cell nucleus showing chromatin, genes, RNA, DNA methylation, DNase I hypersensitive sites, and histone modifications. To the right, there is a "VIEW/DOWNLOAD QUICK LINKS" section with links to "Genome Browsers" (http://www.epigenomebrowser.org, http://genomebrowser.wustl.edu/, http://epigenomegateway.wustl.edu/) and "Data Repositories" (NCBI Epigenomics Gateway, Epigenome Atlas). At the bottom, there is a "NEWS" section with a date "13 DEC" and a link to "Disease Risk Links to Gene Regulation".

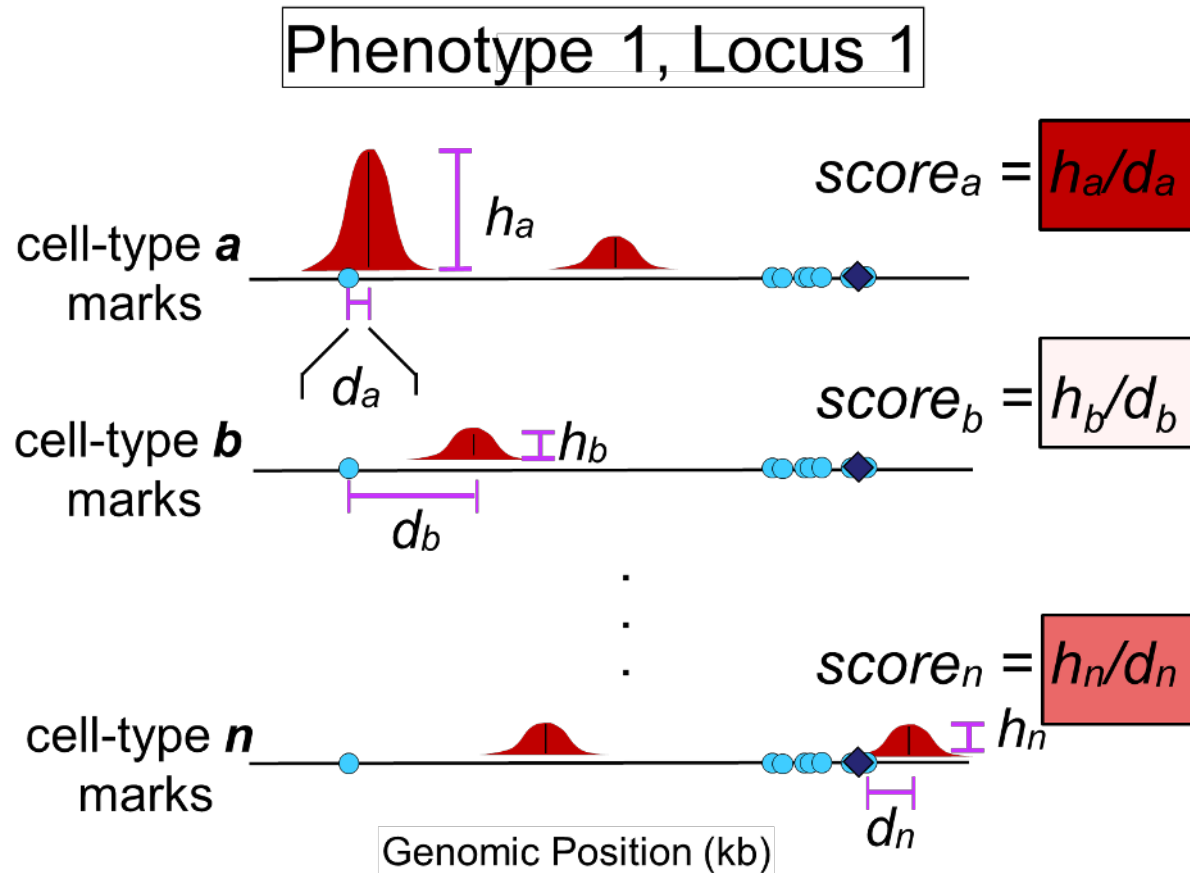
1) Utilise LD information

A. For phenotypically associated variants, find other variants in tight LD ($r^2 > 0.8$)



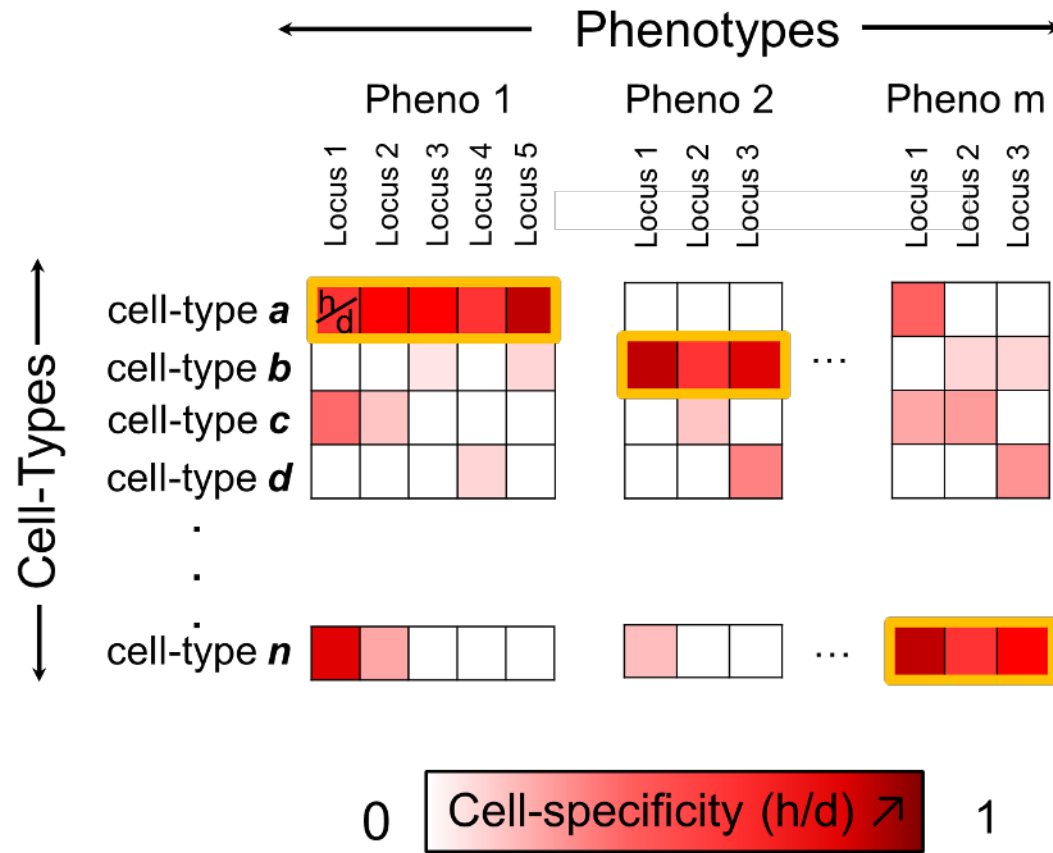
Define best scoring variant at each locus

B. Score each locus based on height and distance of the nearest peak to a variant in tight LD



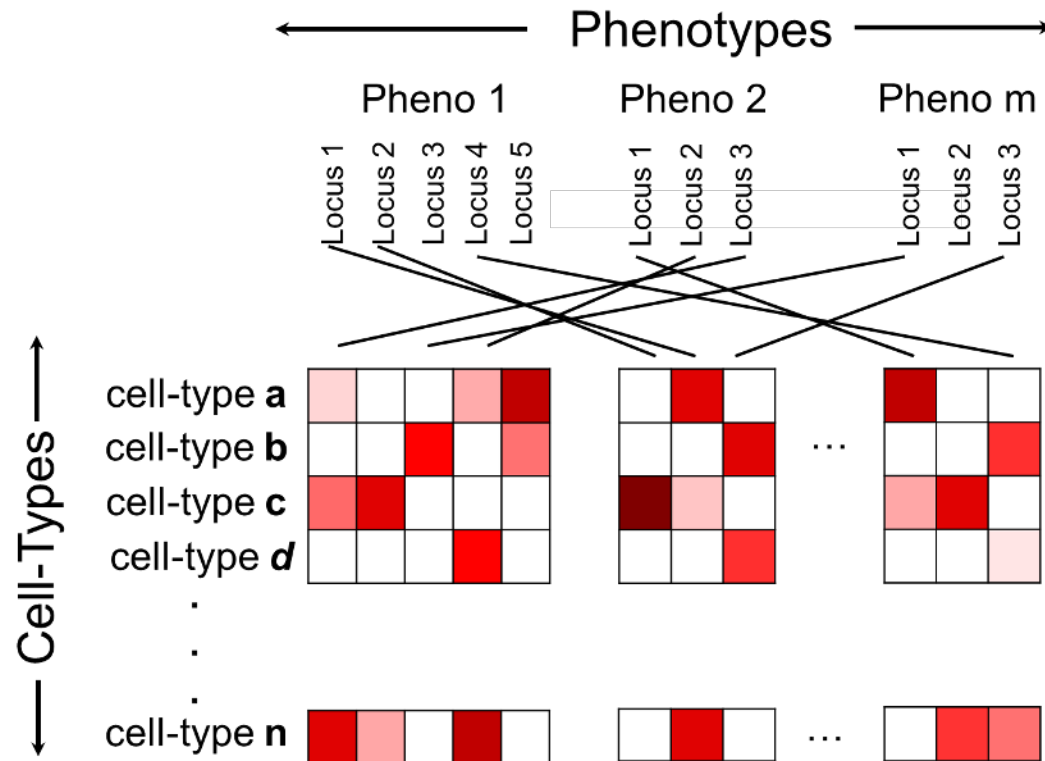
Quantify cell-type specificity of each histone mark

C. Across many phenotypes, assess if marks overlap alleles in specific cell-types

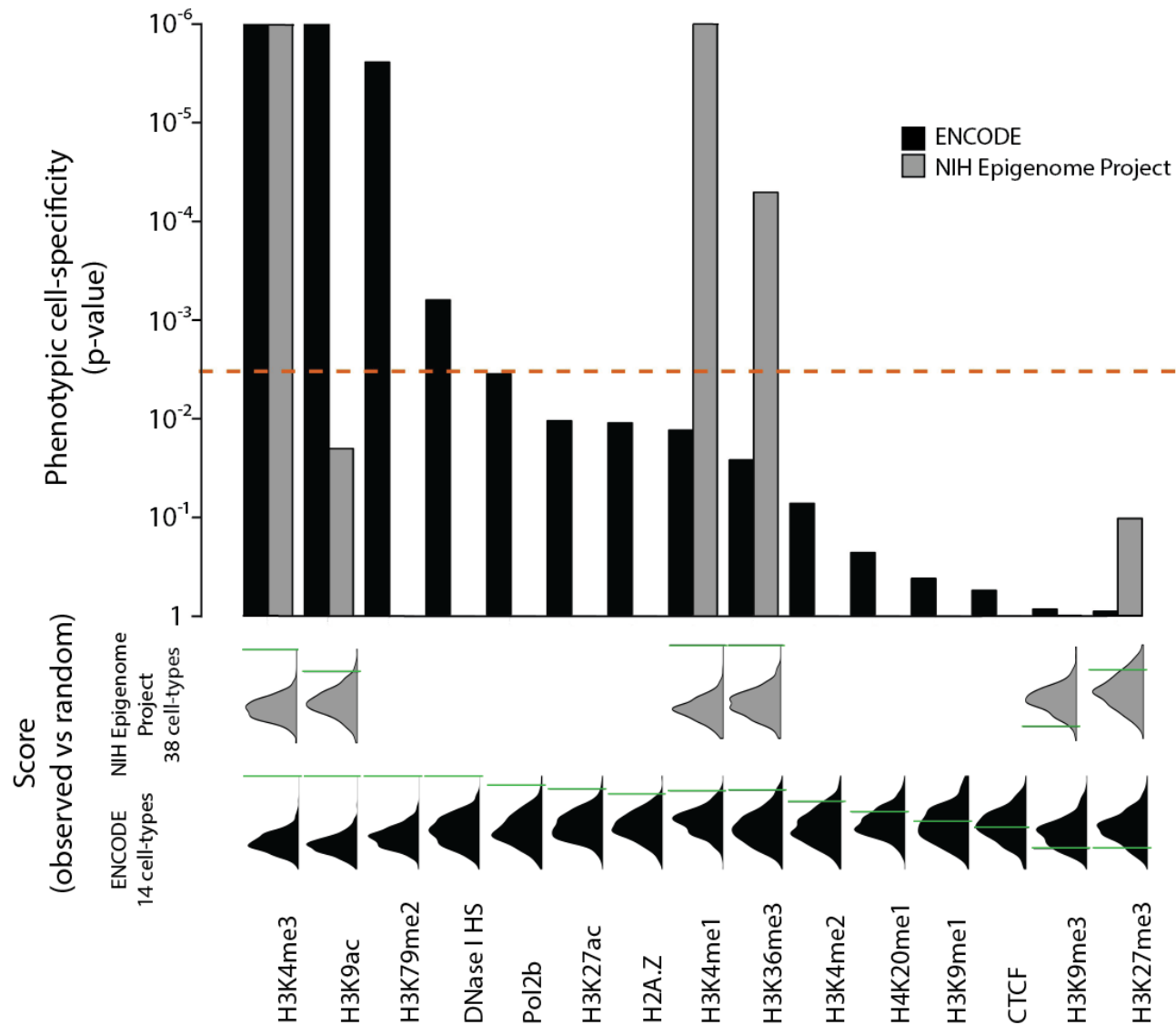


Assess the significance of cell type sepcific signal

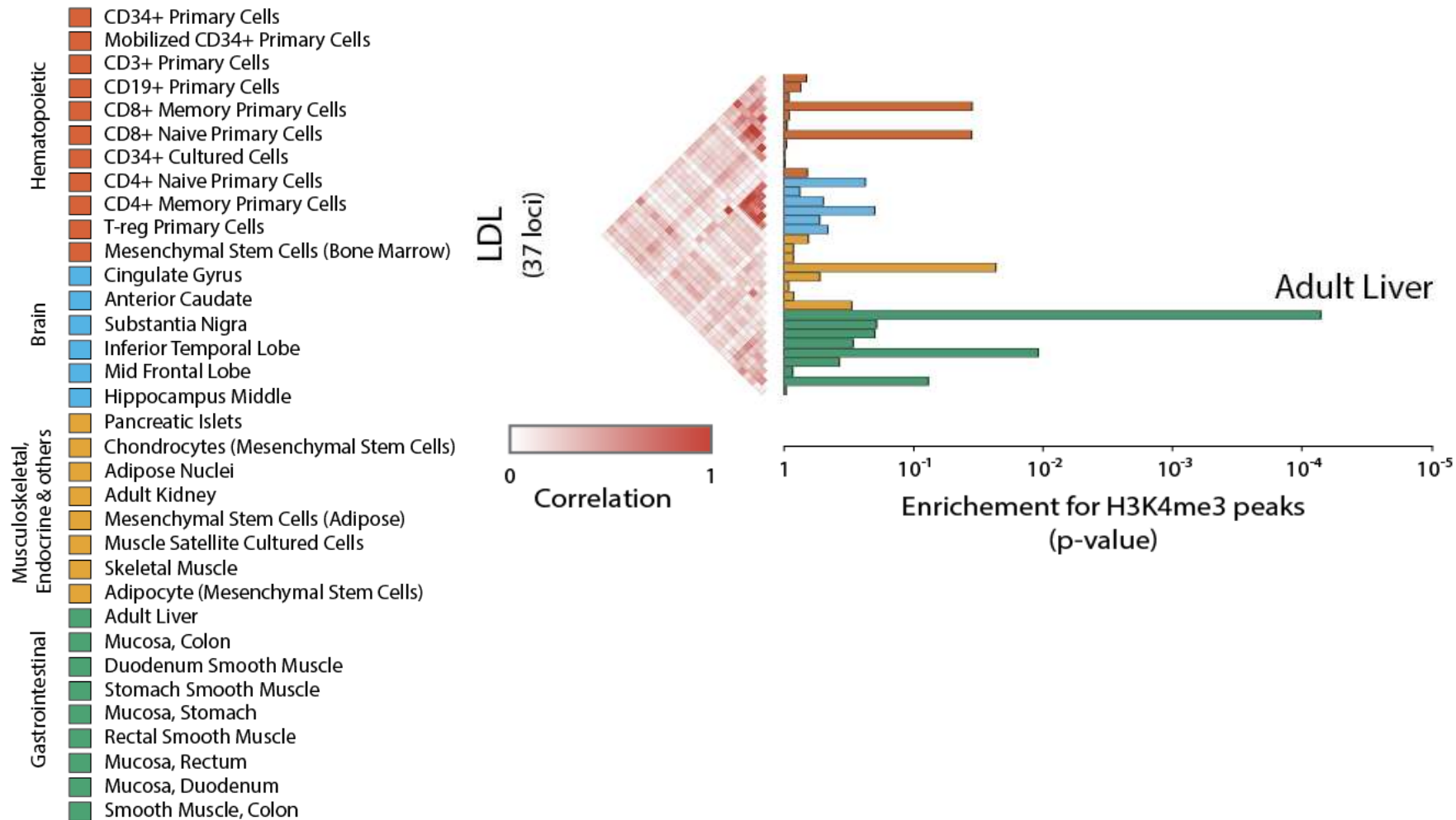
D. Permute to assess significance of phenotypic cell-specificity



Informative marks – active gene regulation

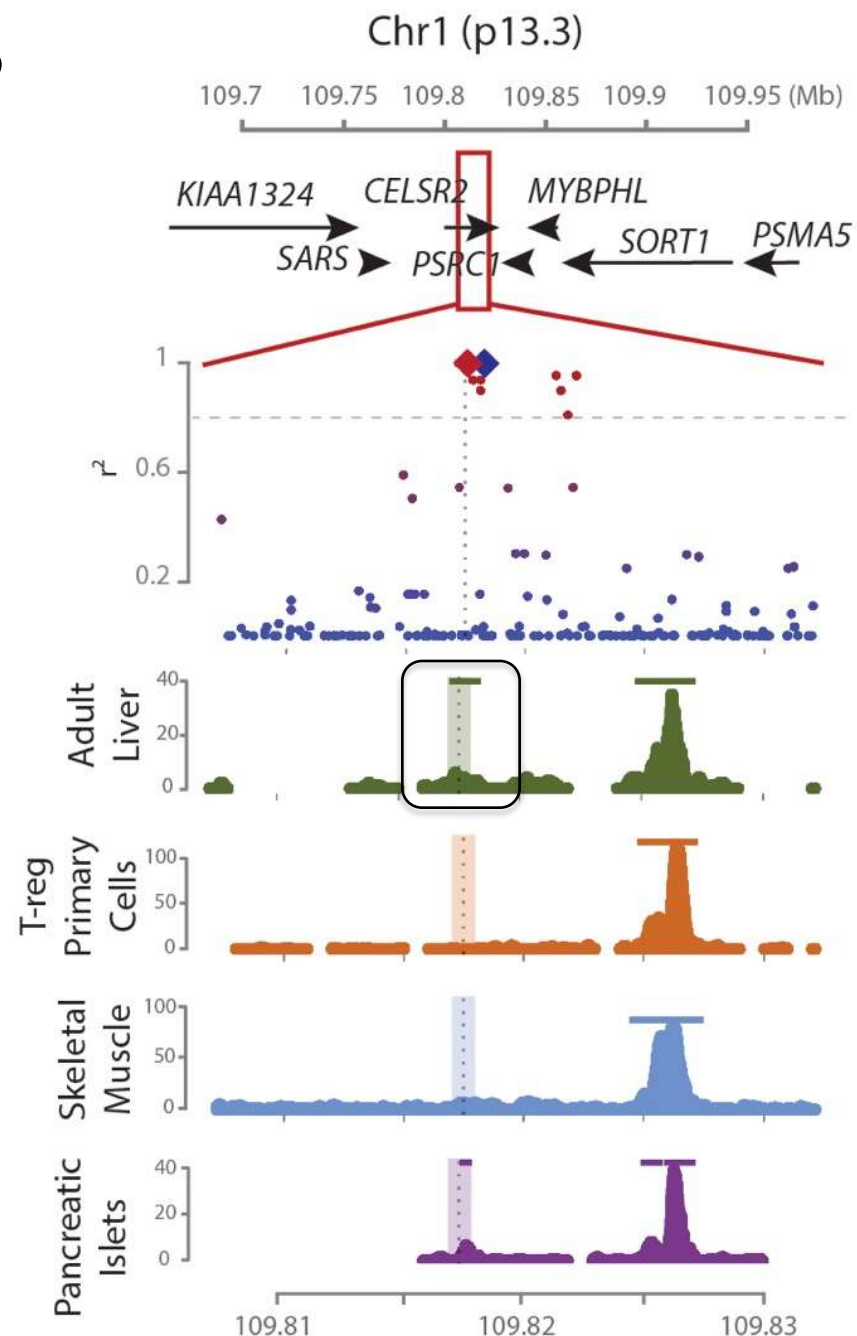


Cell type specific H3K4me3 overlap with LDL SNPs



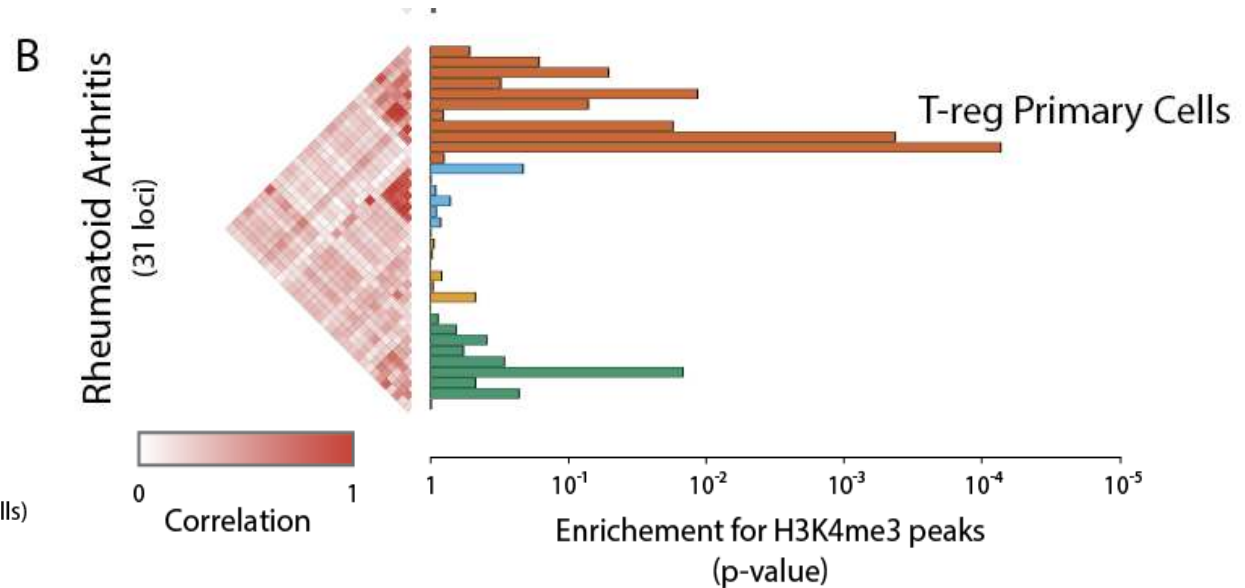
LDL – *SORT1* locus

- SNP closest to H3K4me3 liver specific peak is rs12740374 (87 bp away from the summit)
- Known functional variant! (Musunuru et al. *Nature* 2010)



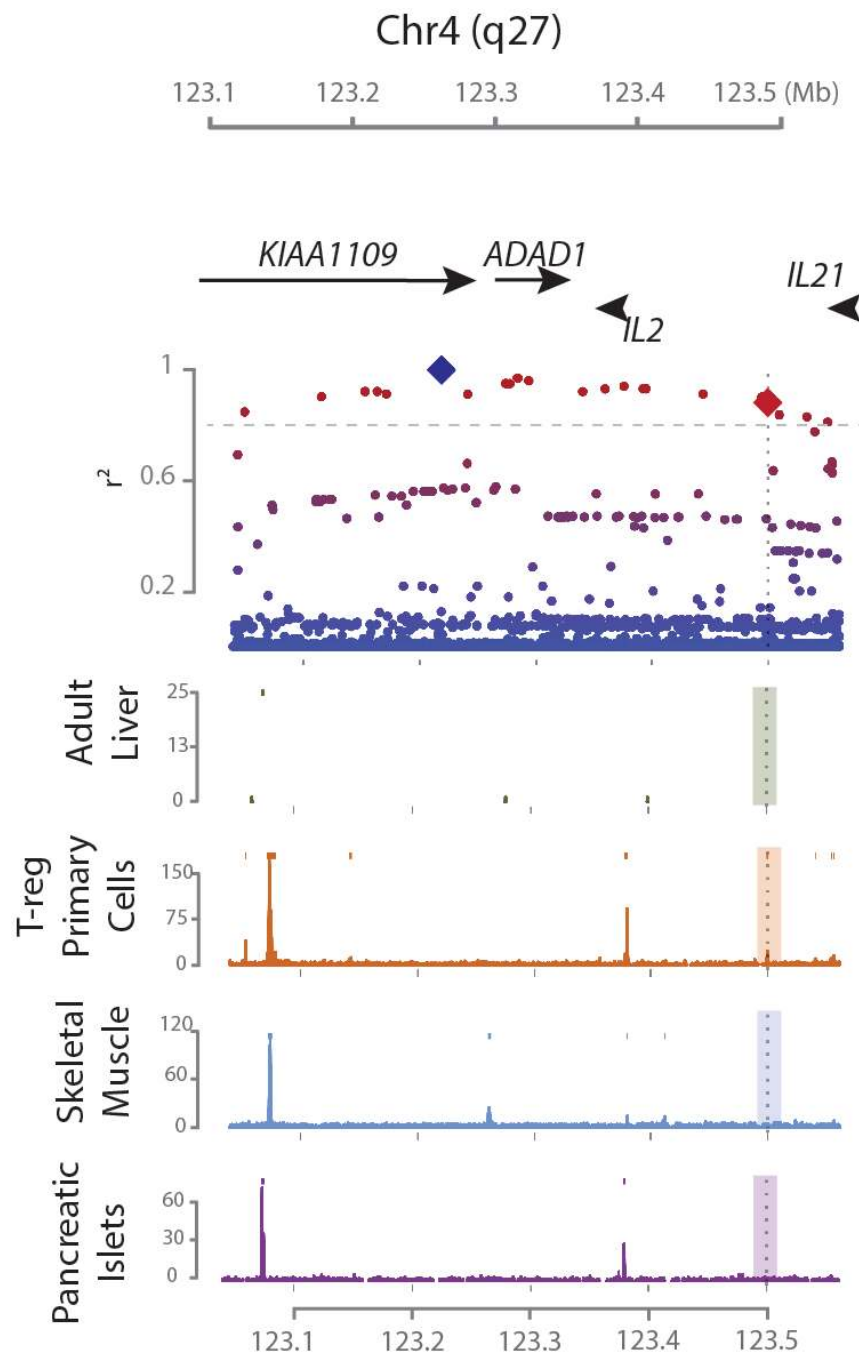
RA: CD4+ cells

- Hematopoietic**
 - CD34+ Primary Cells
 - Mobilized CD34+ Primary Cells
 - CD3+ Primary Cells
 - CD19+ Primary Cells
 - CD8+ Memory Primary Cells
 - CD8+ Naive Primary Cells
 - CD34+ Cultured Cells
 - CD4+ Naive Primary Cells
 - CD4+ Memory Primary Cells
 - T-reg Primary Cells
 - Mesenchymal Stem Cells (Bone Marr)
- Brain**
 - Cingulate Gyrus
 - Anterior Caudate
 - Substantia Nigra
 - Inferior Temporal Lobe
 - Mid Frontal Lobe
 - Hippocampus Middle
- Musculoskeletal, Endocrine & others**
 - Pancreatic Islets
 - Chondrocytes (Mesenchymal Stem Cells)
 - Adipose Nuclei
 - Adult Kidney
 - Mesenchymal Stem Cells (Adipose)
 - Muscle Satellite Cultured Cells
 - Skeletal Muscle
 - Adipocyte (Mesenchymal Stem Cells)
- Gastrointestinal**
 - Adult Liver
 - Mucosa, Colon
 - Duodenum Smooth Muscle
 - Stomach Smooth Muscle
 - Mucosa, Stomach
 - Rectal Smooth Muscle
 - Mucosa, Rectum
 - Mucosa, Duodenum
 - Smooth Muscle, Colon



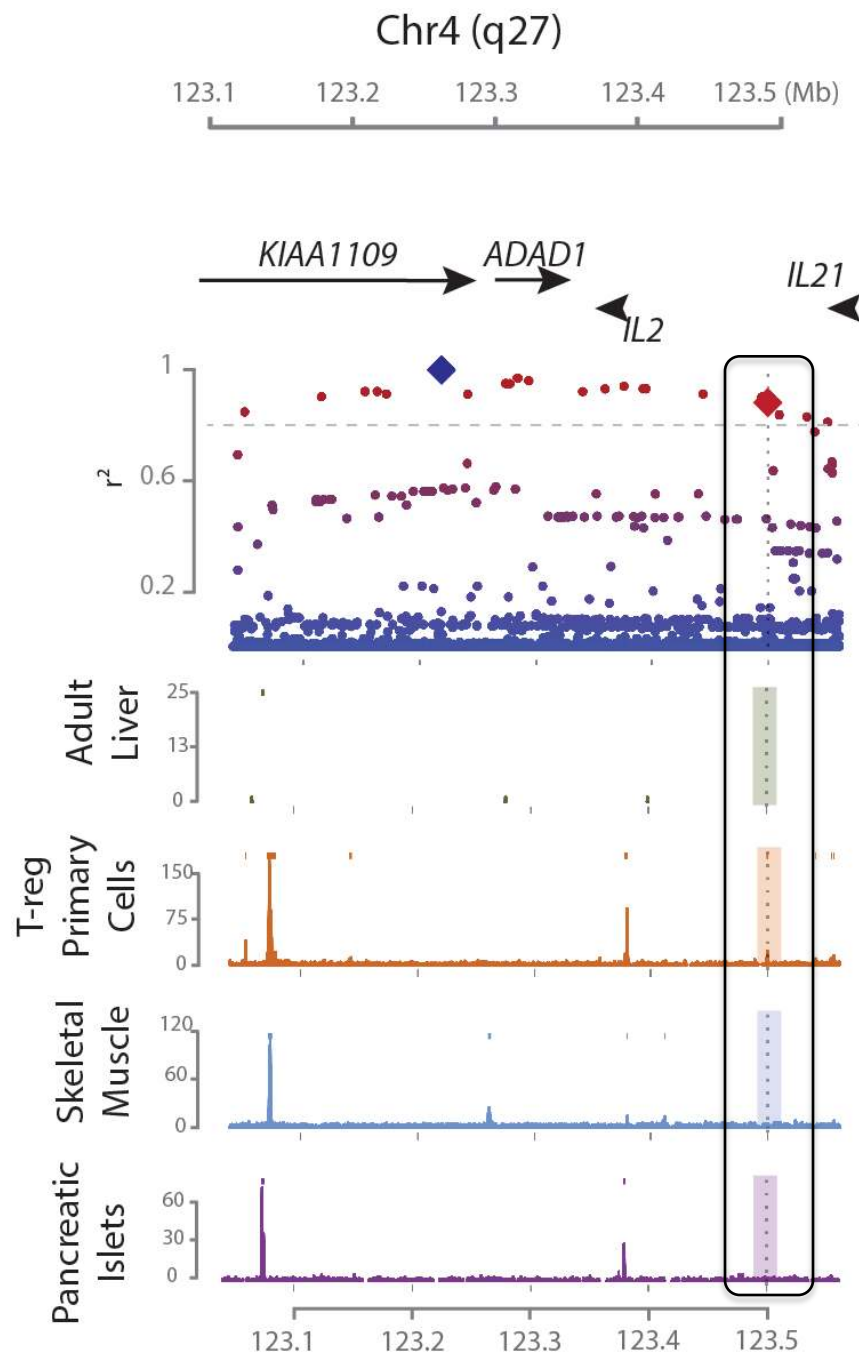
RA – *IL2/IL21* locus

- Reported associated rs13119723 SNP is within *KIAA1109*
- LD block of over 500-kb
- rs13140464, 116 bp from the summit of the peak highly specific to the T-reg cells, located between *IL2* and *IL21* genes



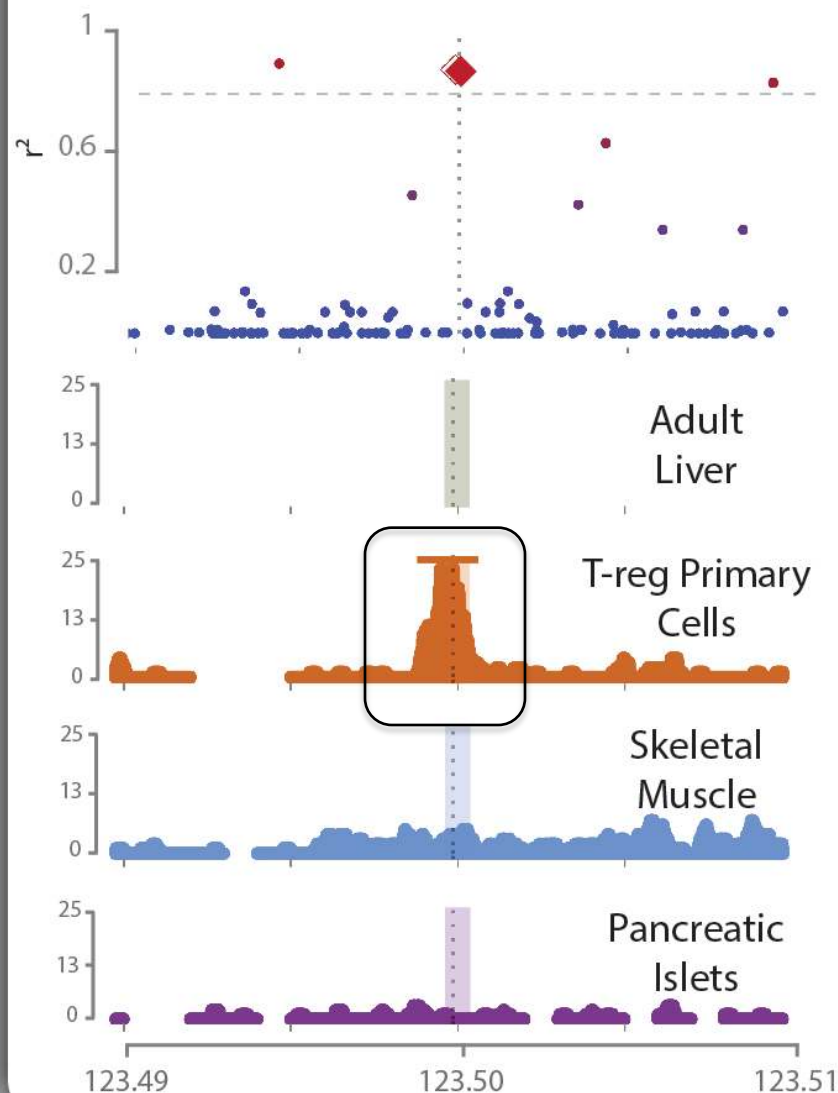
RA – *IL2/IL21* locus

- Reported associated rs13119723 SNP is within *KIAA1109*
- LD block of over 500-kb
- rs13140464, 116 bp from the summit of the peak highly specific to the T-reg cells, located between *IL2* and *IL21* genes

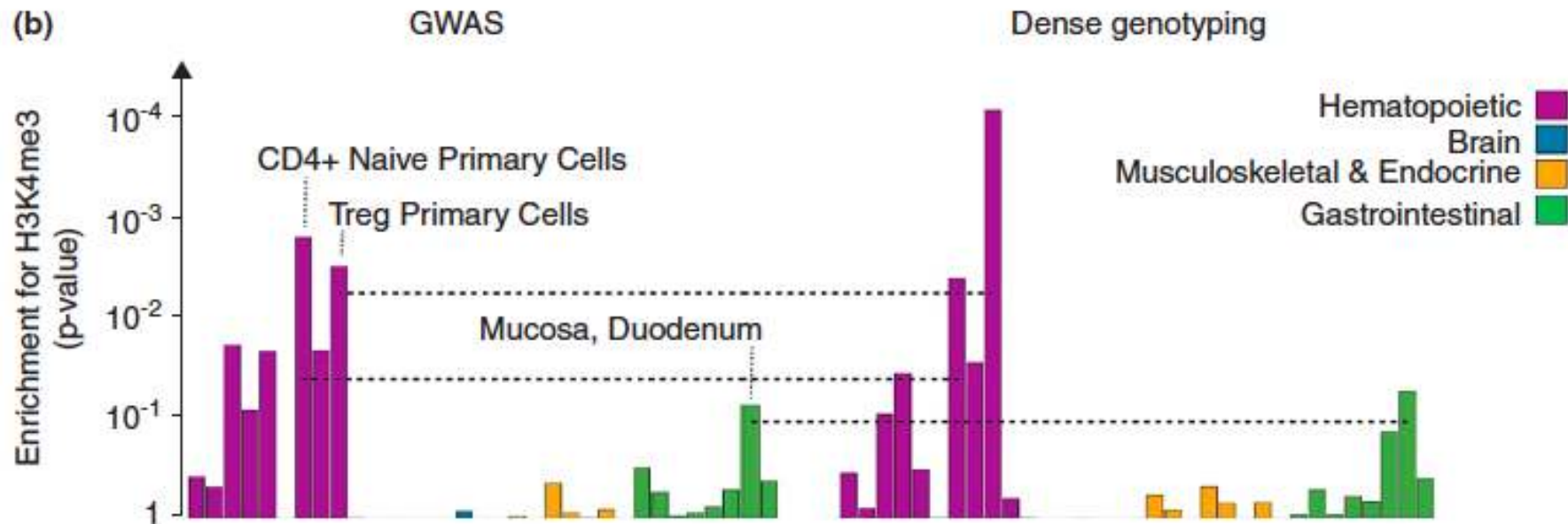


RA – *IL2/IL21* locus

- Reported associated rs13119723 SNP is within *KIAA1109*
- LD block of over 500-kb
- rs13140464, 116 bp from the summit of the peak highly specific to the T-reg cells, located between *IL2* and *IL21* genes

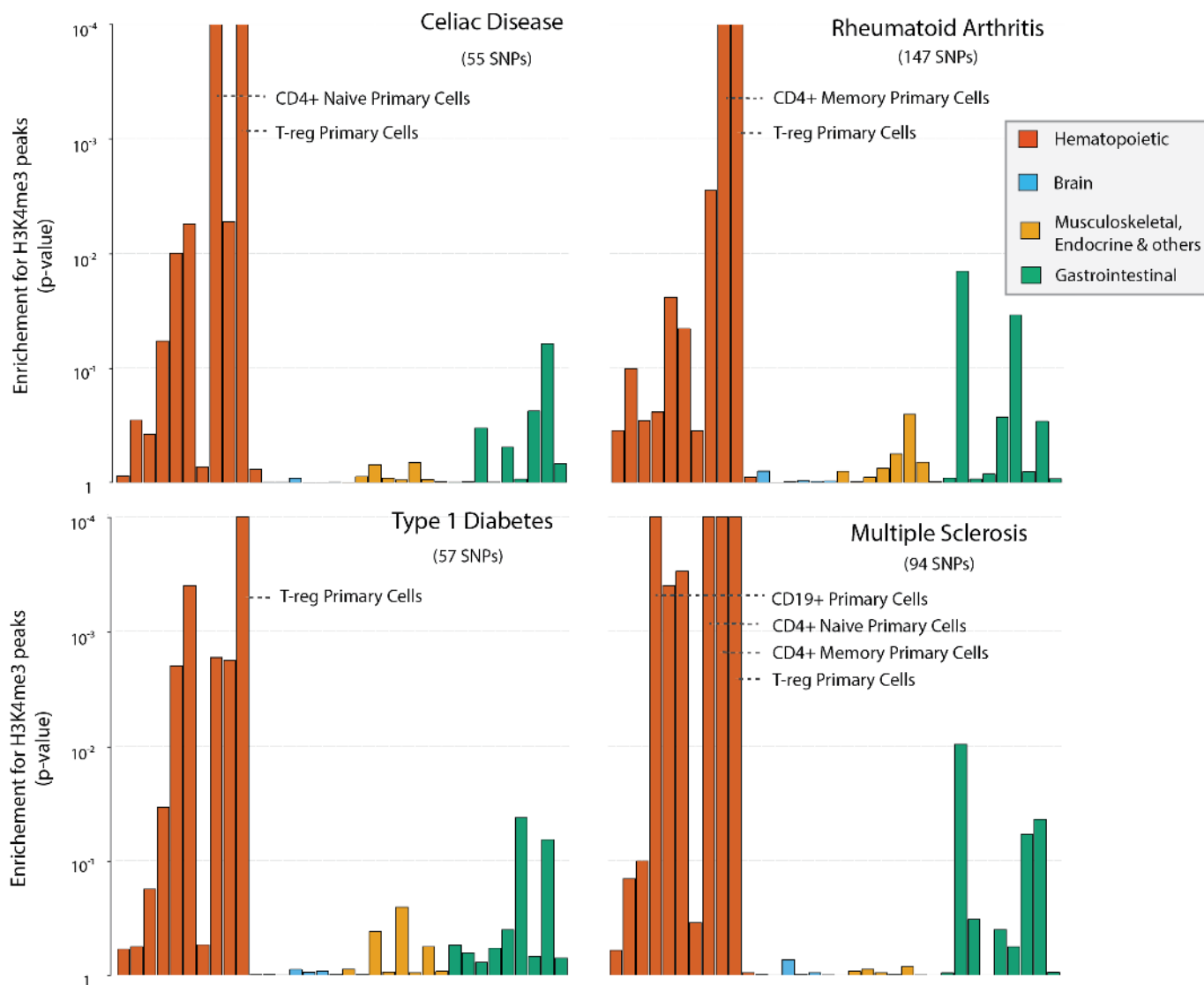


Statistical refinement of association signals improves tissue specific signal

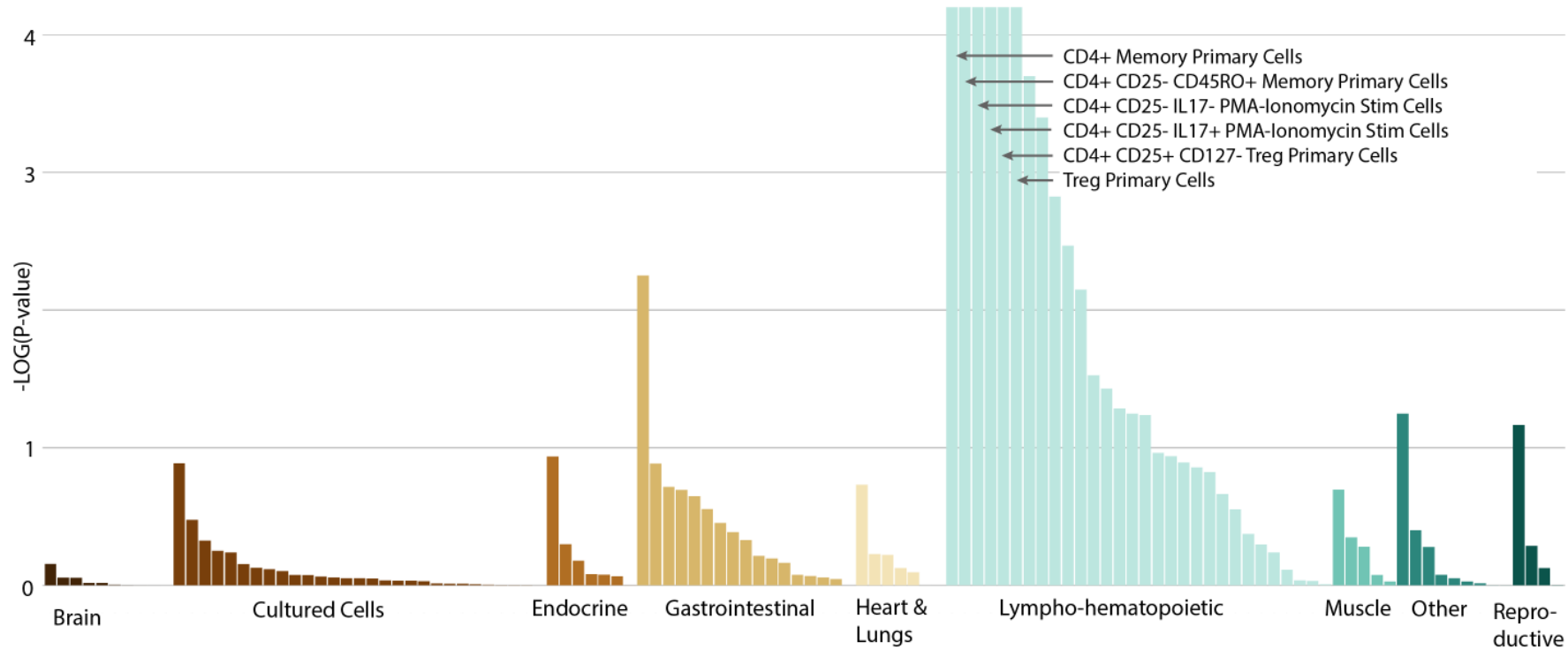


Trynka and Raychaudhuri, *Curr Opin Genet Dev* 2013

The role of CD4+ subsets in autoimmune diseases



T1D variants implicate a limited set of cell types



Predicting Cell Types and Genetic Variations Contributing to Disease by Combining GWAS and Epigenetic Data

Anna Gerasimova*, Lukas Chavez, Bin Li, Gregory Seumois, Jason Greenbaum, Anjana Rao, Pandurangan Vijayanand, Bjoern Peters

La Jolla Institute for Allergy and Immunology, La Jolla, California, United States of America

Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression

Richard Cowper-Sallari^{1,2,7}, Xiaoyang Zhang^{1,2,7}, Jason B Wright¹, Swneke D Bailey^{3,4}, Jerome Eeckhoutte^{5,6}, Jason H Moore^{1,2} & Mathieu Lupien^{3,4}

Maps of Open Chromatin Guide the Functional Follow-Up of Genome-Wide Association Signals: Application to Hematological Traits

Dick S. Paul^{1*}, James B. Nicholas¹, Tony De Vries¹, Stuart Maccham^{1,2,3}, Augusto Rendon^{2,3,4}, Katta Venkatesan^{2,5}, Suthesh Sivapalaratnam⁶, Hanneke Huisman¹, Berthold Göttgens^{2,5}, Nicole...

Systematic functional regulatory assessment of disease-associated variants

Konrad J. Karczewski^{a,b,1}, Joel T. Dudley^{a,c,1,2}, Kimberly R. Kukurba^b, Rong Chen^c, Atul J. Butte^c, Stephen B. Montgomery^{b,d}, and Michael Snyder^{b,3}

Linking disease associations with regulatory information in the human genome

Marc A. Schaub,¹ Alan P. Boyle,² Anshul Kundaje,¹ Serafim Batz...

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

Matthew T. Maurano,^{1*} Richard Humbert,^{1*} Eric Rynes,^{1*} Robert E. Thurman,¹ Eric Haugen,¹ Hao Wang,¹ Alex P. Reynolds,¹ Richard Sandstrom,¹ Hongzhu Qu,^{1,2} Jennifer Brody,³ Anthony Shafer,¹ Fidencio Neri,¹ Kristen Lee,¹ Tanya Kutayavin,¹ Sandra Stehling-Sun,¹ Audra K. Johnson,¹ Theresa K. Canfield,¹ Erika Giste,¹ Morgan Diegel,¹ Daniel Bates,¹ R. Scott Hansen,⁴ Shane Neph,¹ Peter J. Sabo,¹ Shelly Heimfeld,⁵ Antony Raubitschek,⁶ Steven Ziegler,⁶ Chris Cotsapas,^{7,8} Nona Sotoodehnia,^{3,9} Ian Glass,¹⁰ Shamil R. Sunyaev,¹¹ Rajinder Kaul,⁴ John A. Stamatoyannopoulos^{1,12†}

ARTICLE

doi:10.1038/

An integrated encyclopedia of DNA elements in the human genome

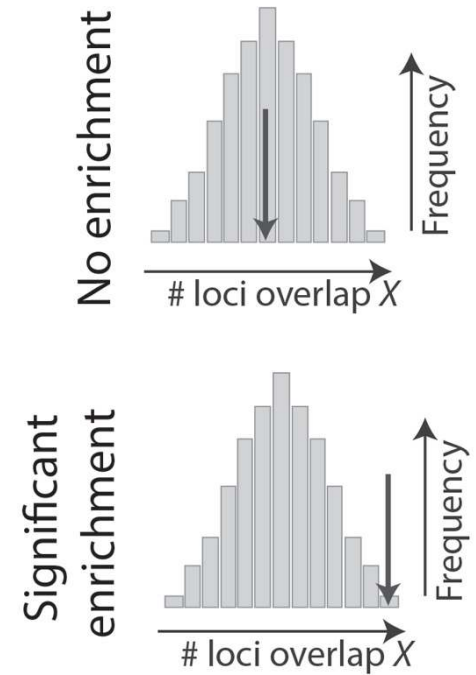
The ENCODE Project Consortium*

GWAS variants and enrichment across a range of genomic annotations

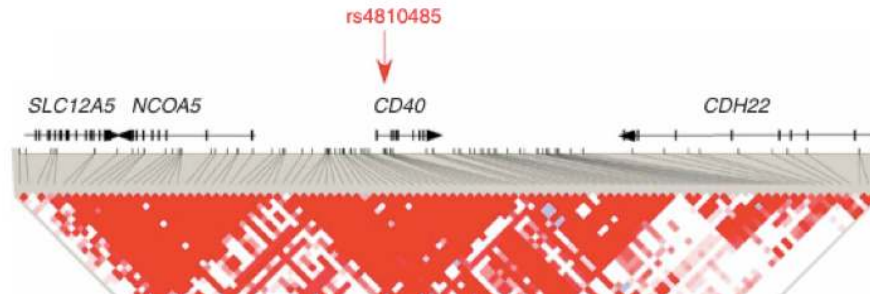
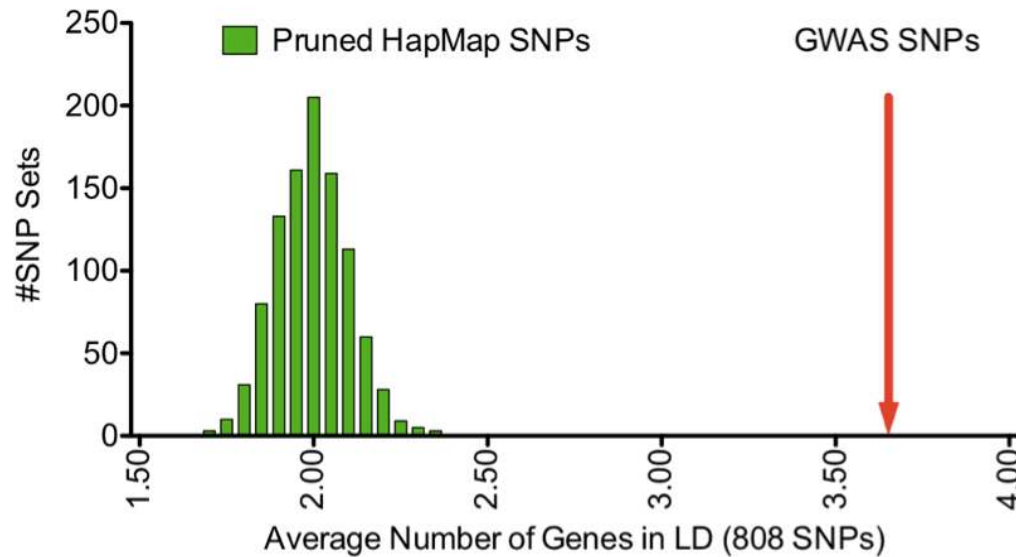
- A robust test to assess enrichment of disease SNPs with any genomic annotation

Caveats in enrichment testing

- Matching based enrichment tests can be biased – require a prior knowledge of relevant matching parameters

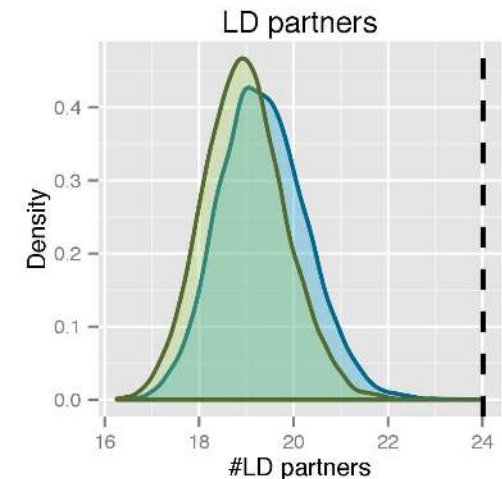
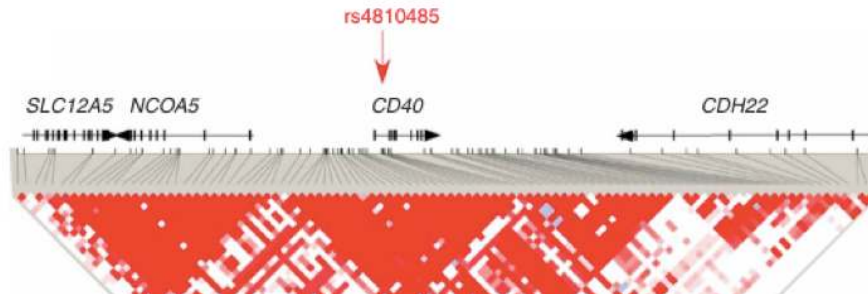
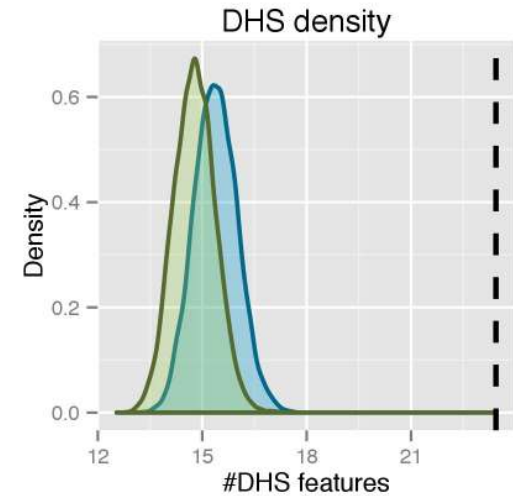
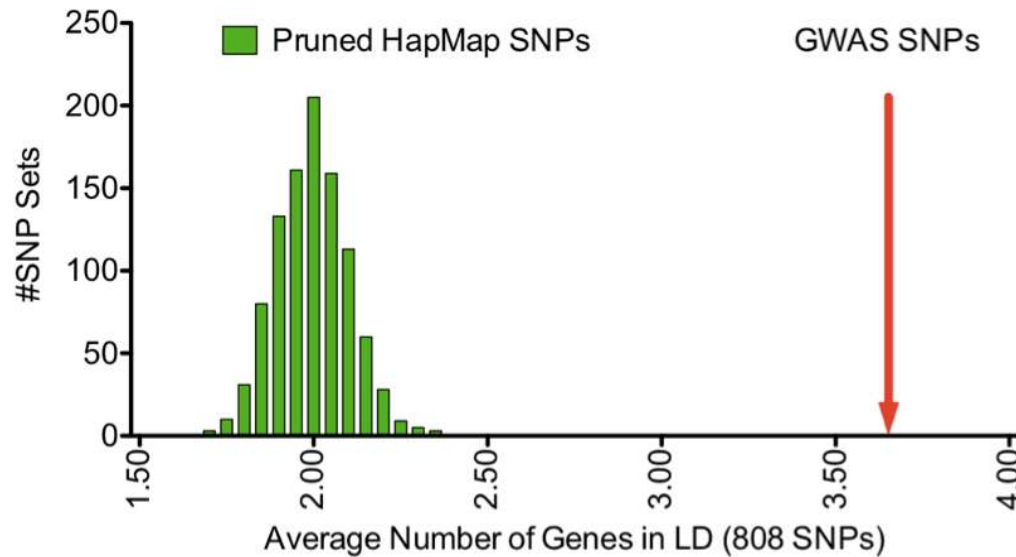


Distinct properties at GWAS associated loci

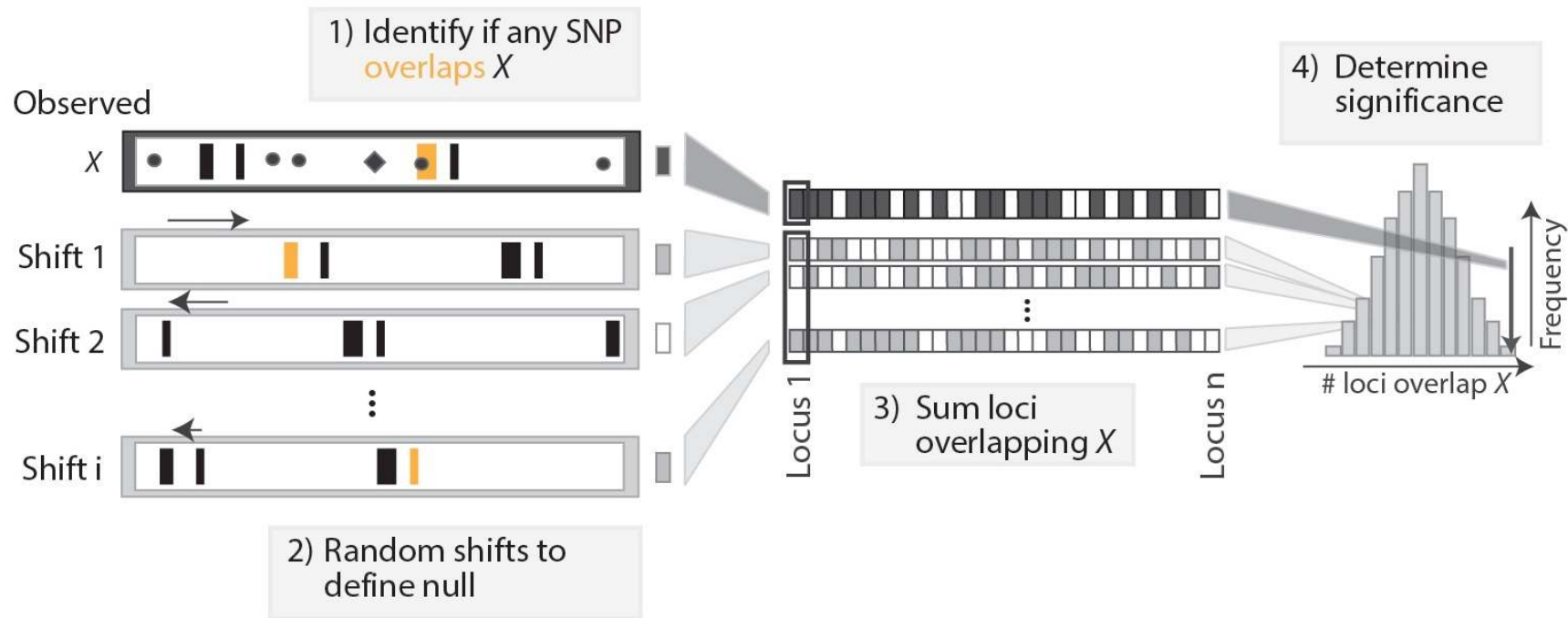


Distinct properties at GWAS associated loci

Matched
Random

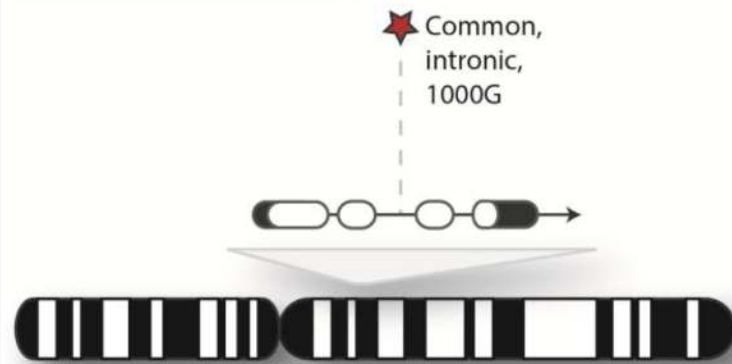


Genomic Annotation Shifter (*GoShifter*)

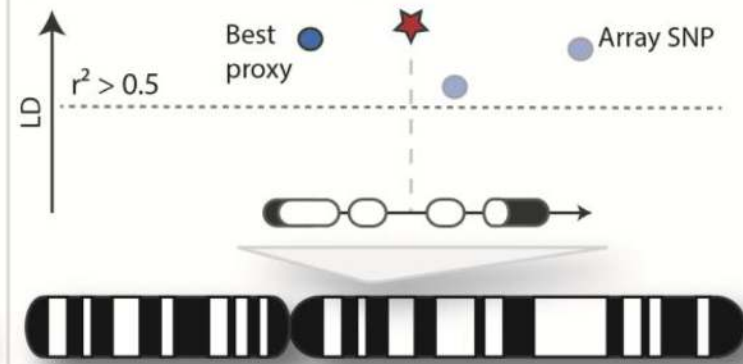


Simulate GWAS results

1) Define causal variant

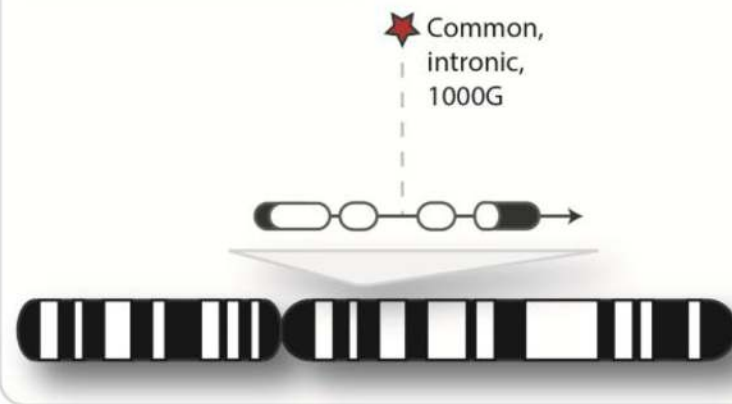


2) Find best array proxy

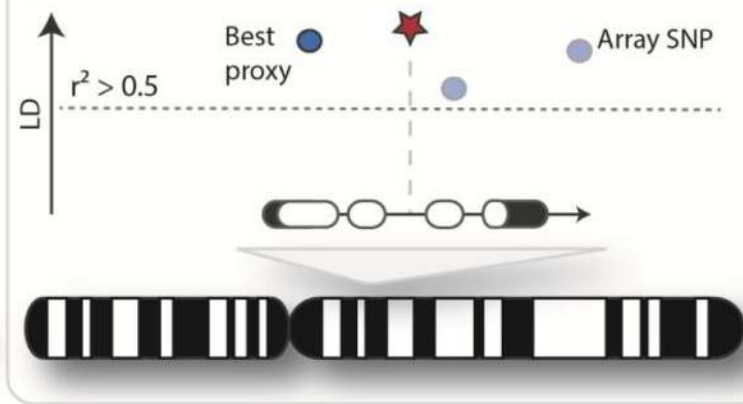


Simulate GWAS results

1) Define causal variant

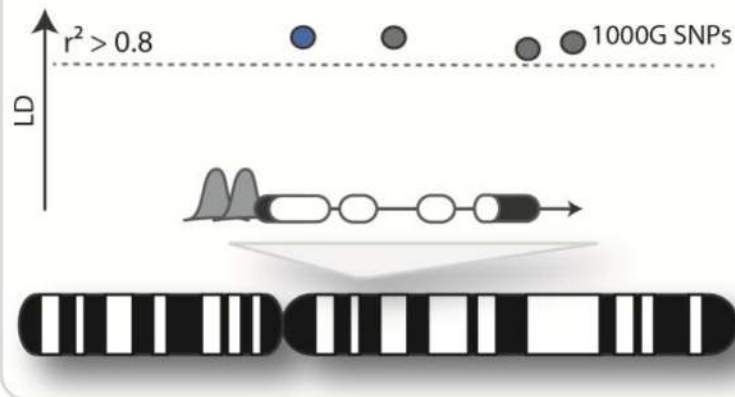


2) Find best array proxy

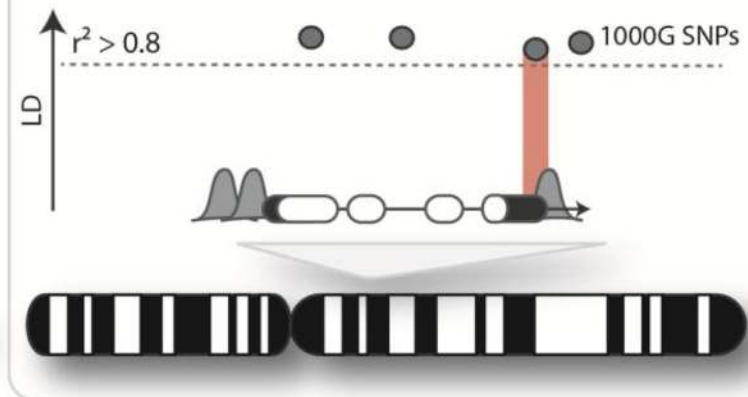


Test for enrichment

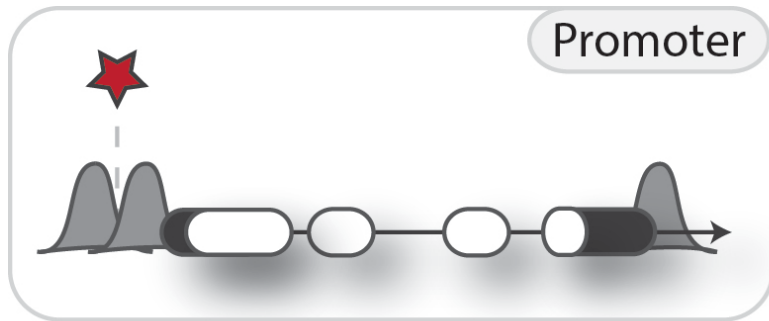
3) Identify LD partners



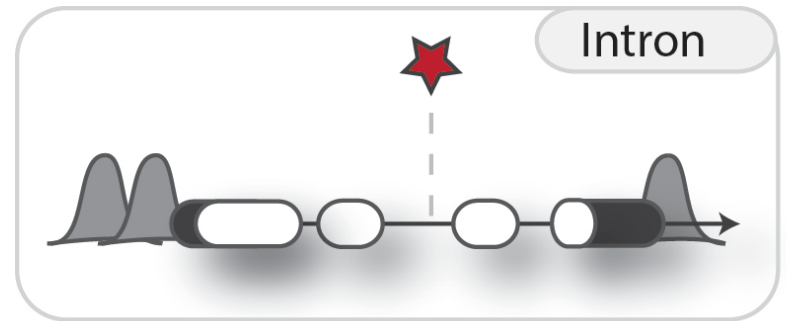
3) Test for overlap with annotation



Assumption



- Promoter catalogs should enrich for DHS peaks

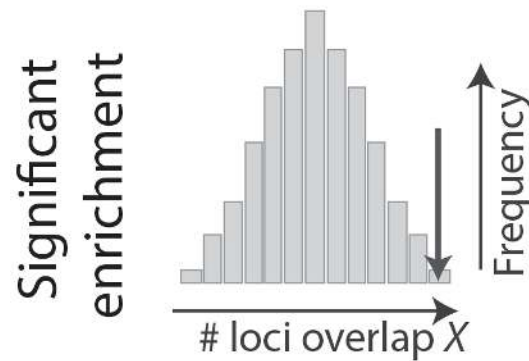
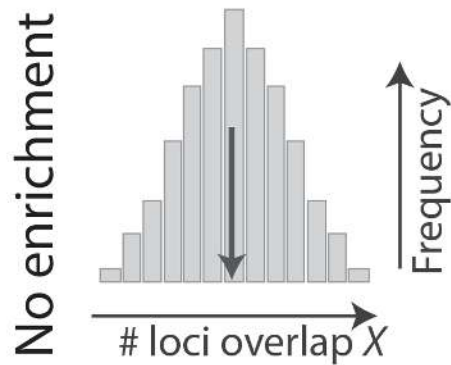


- Intron catalogs should not enrich for DHS peaks

Dataset

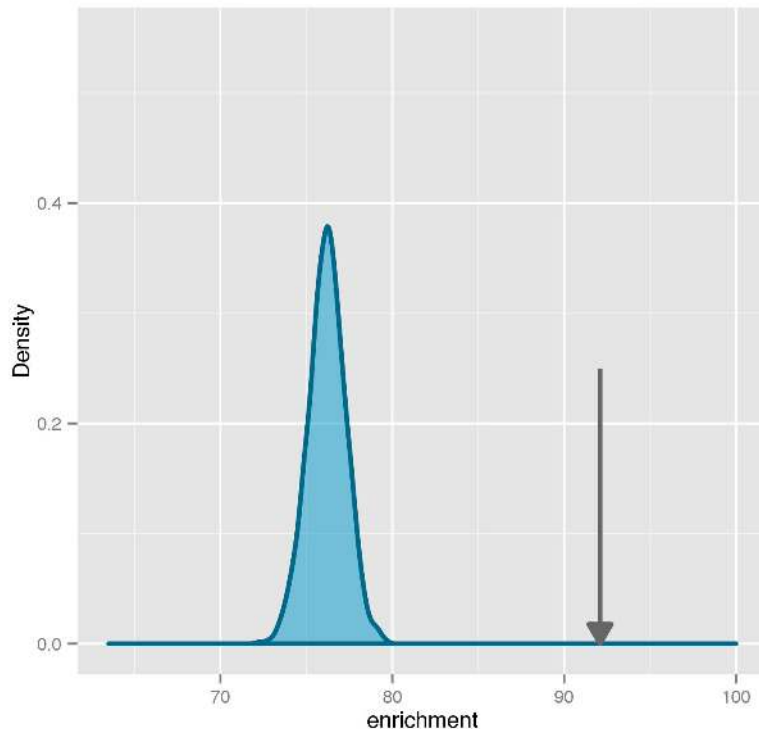
- DHS peaks from the ENCODE and Roadmap Epigenomics Projects
- DNase hypersensitivity data from:
 - 217 cell and tissue samples
 - 1,376,301 distinct DHS positions
 - Collectively spanning 16.4% of the genome

Interpretation of the results



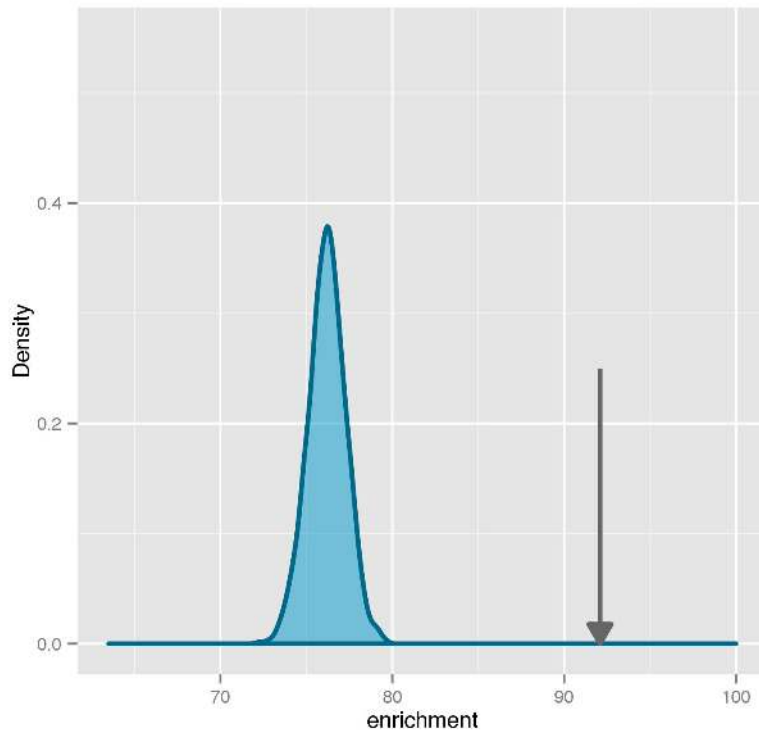
Standard matching on GEN, MAF and TSS

Promoter

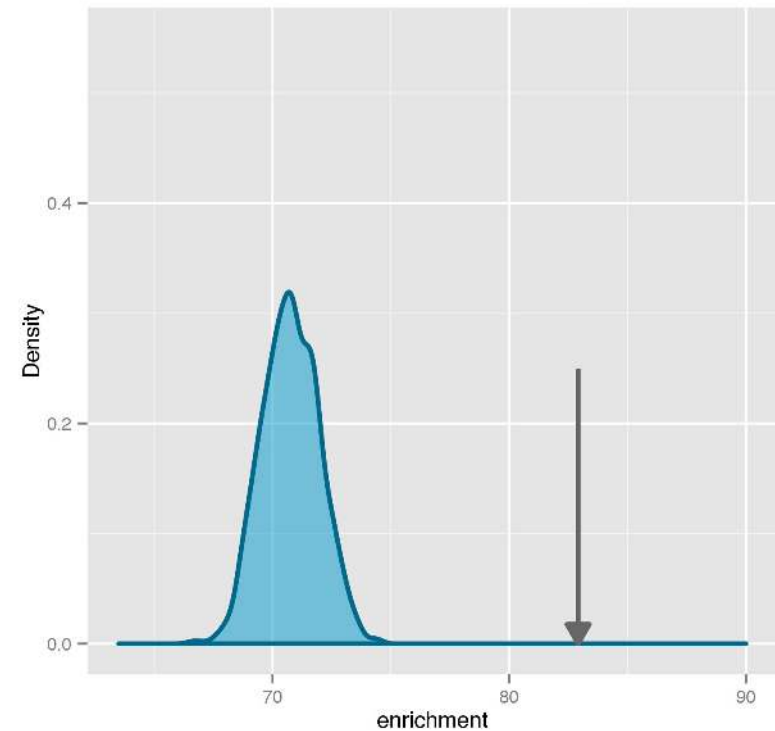


Standard matching on GEN, MAF and TSS

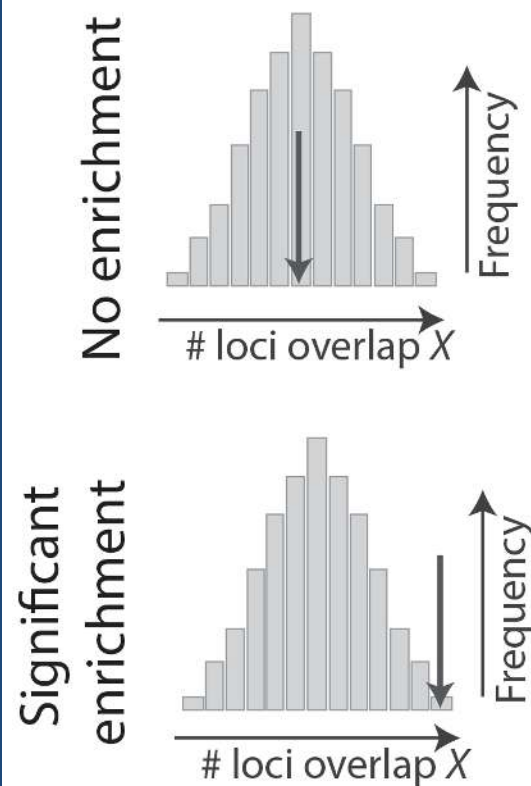
Promoter



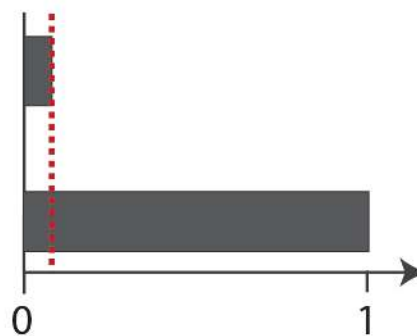
Intron



Interpretation of the results

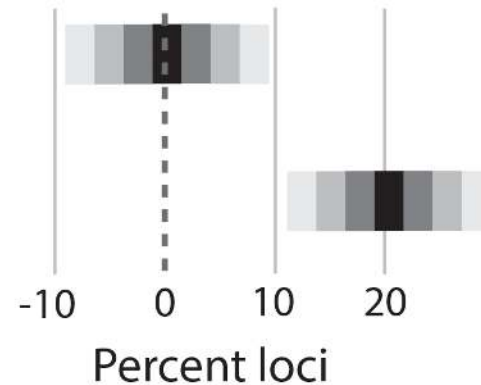


Proportion of
significant results
(in 1000 simulations)



Delta-overlap

Deviation from observed
(observed - expected)



Benchmarking enrichment methods using simulated sets of SNPs

Enrichment test



DHS

Promoter

5'UTR

Functional model

0.00 0.25 0.50 0.75 1.00

Proportion of
significant SNP sets

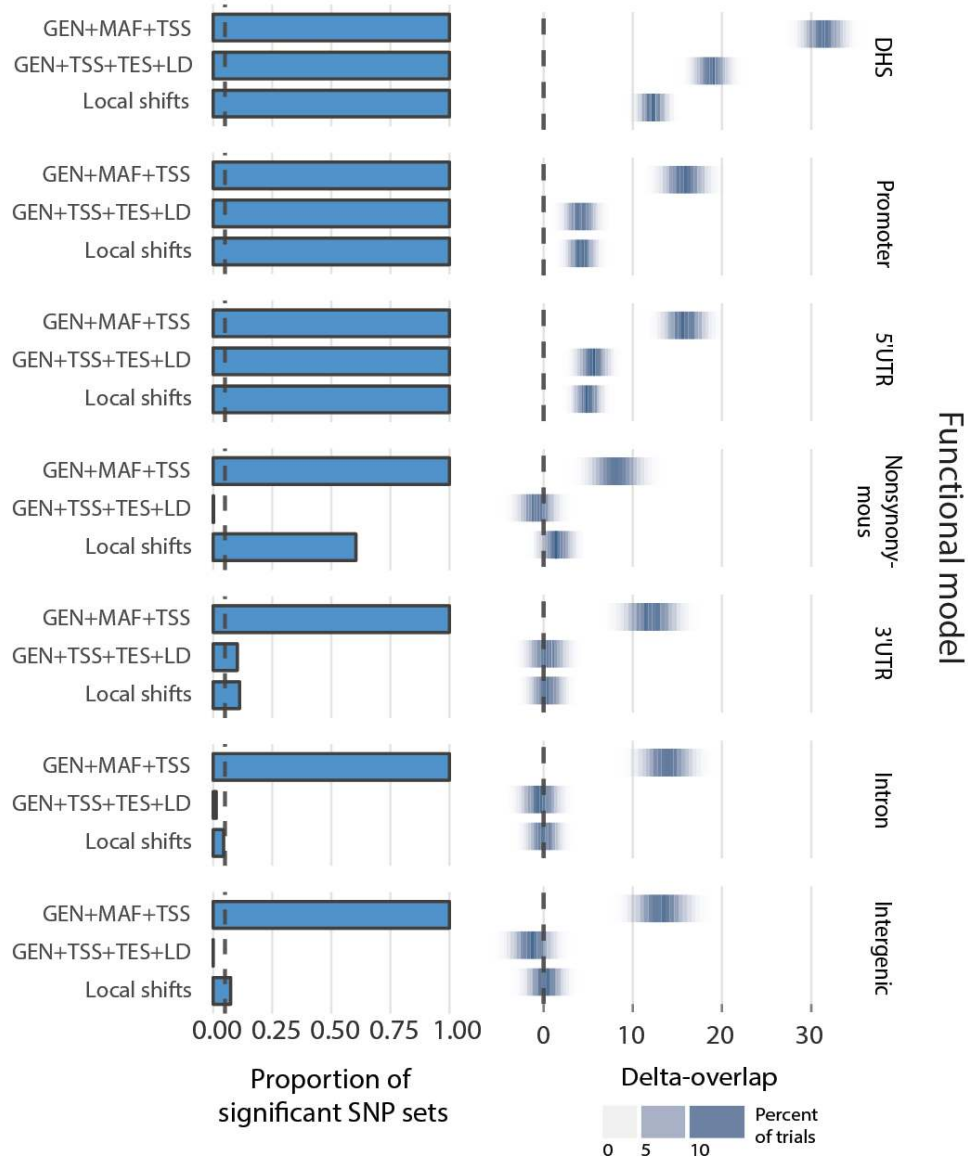
0 10 20 30

Delta-overlap

0 5 10 Percent
of trials

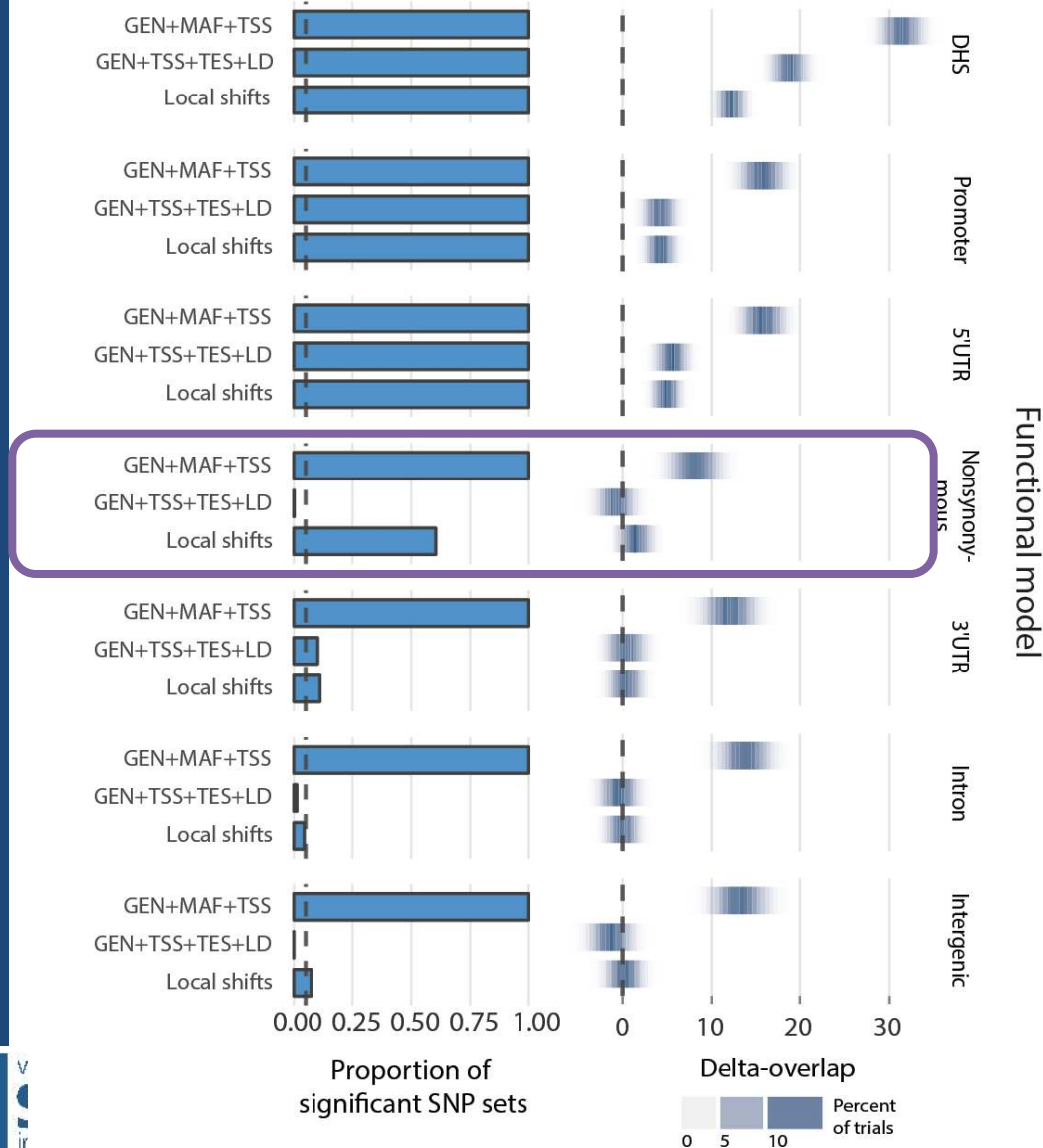
Benchmarking enrichment methods using simulated sets of SNPs

Enrichment test



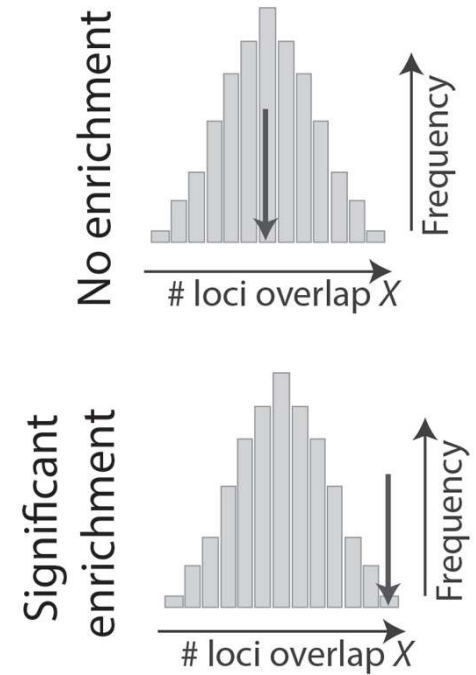
Benchmarking enrichment methods using simulated sets of SNPs

Enrichment test



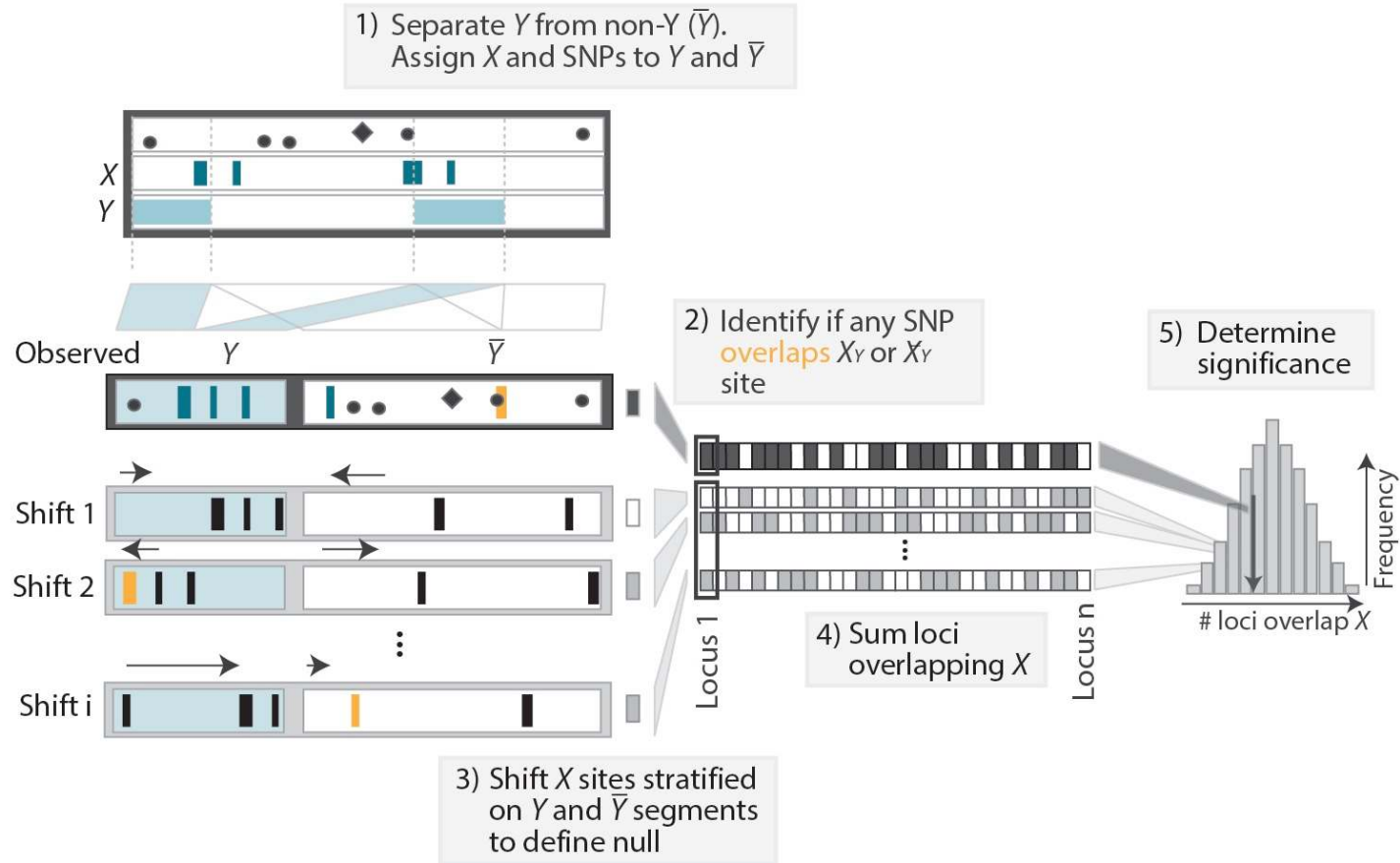
Caveats in enrichment testing

- Matching based enrichment tests can be biased – require a prior knowledge of relevant matching parameters
- Genomic annotations colocalize

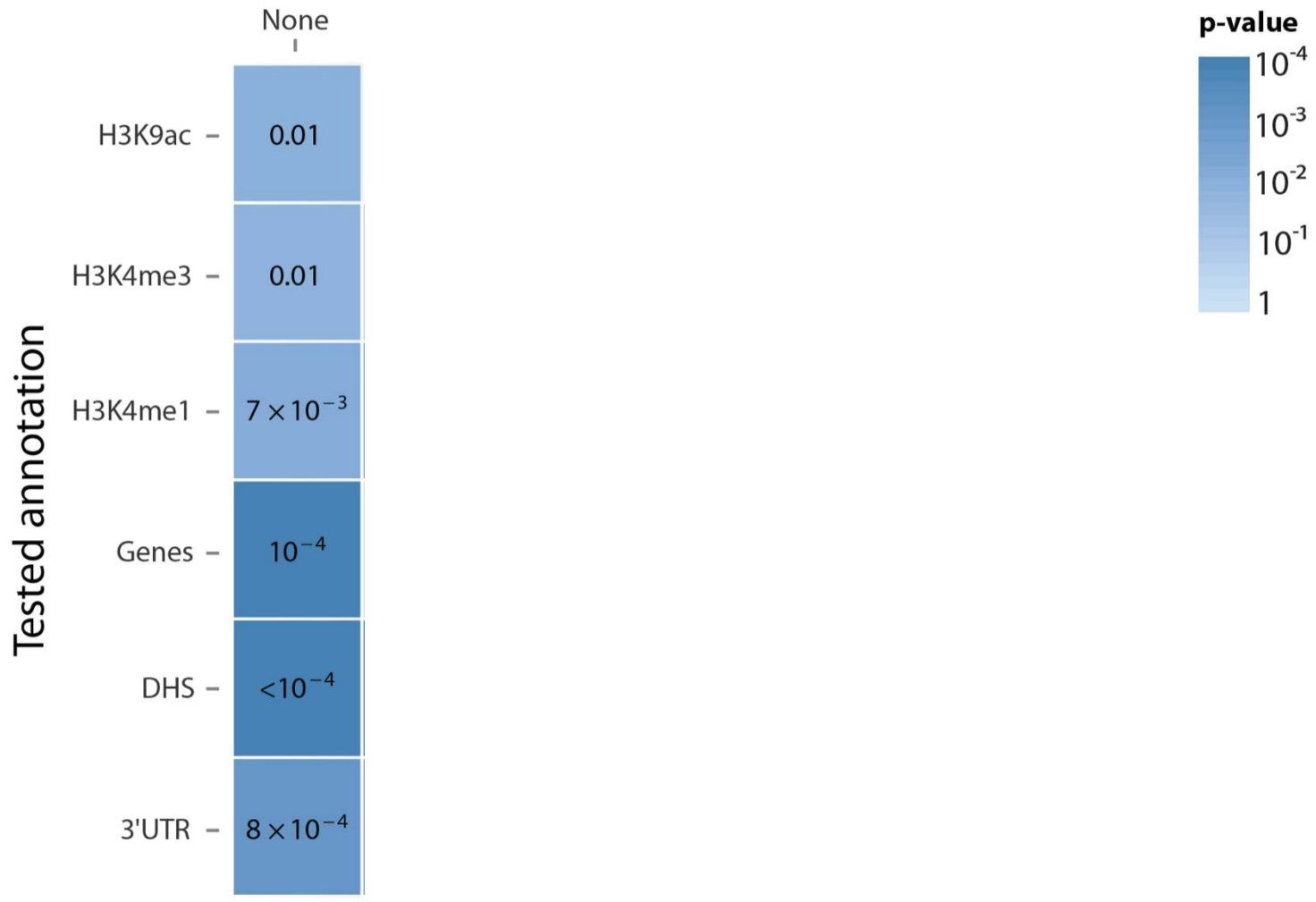


Differentiating effects of colocalizing annotations

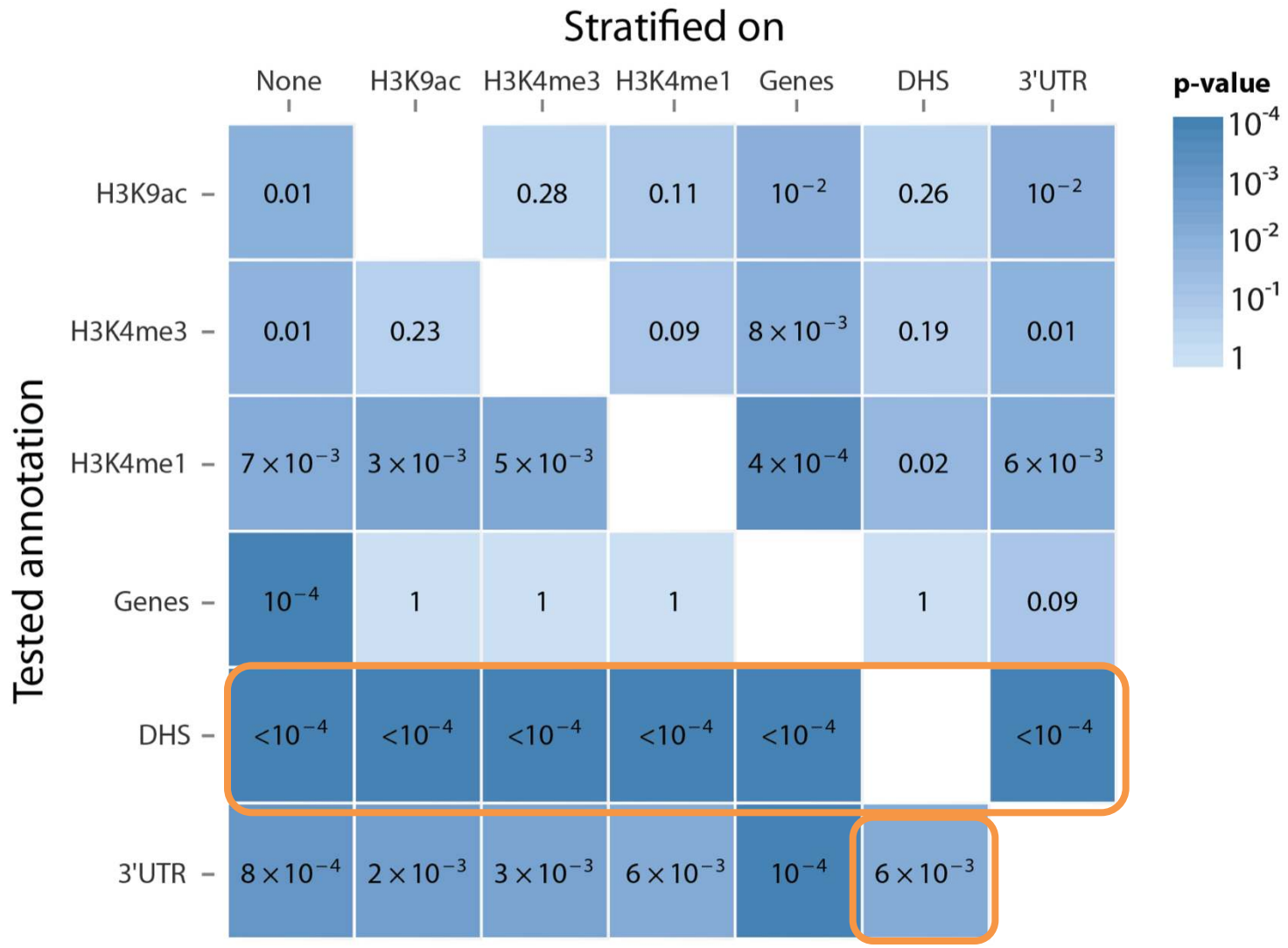
C) Test significance of overlap with annotation X conditioning on Y



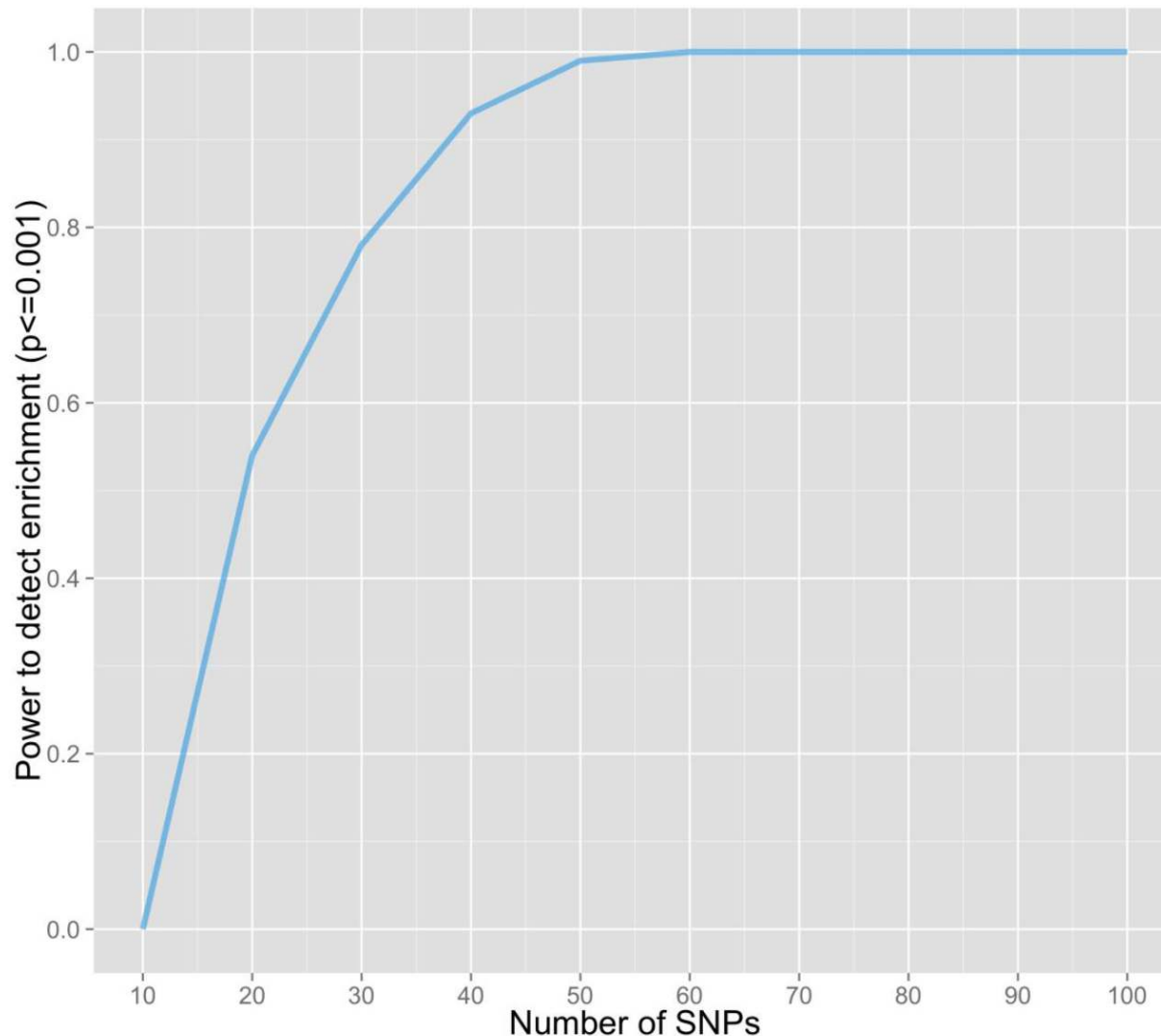
eQTLs are independently enriched for DHS and 3'UTR



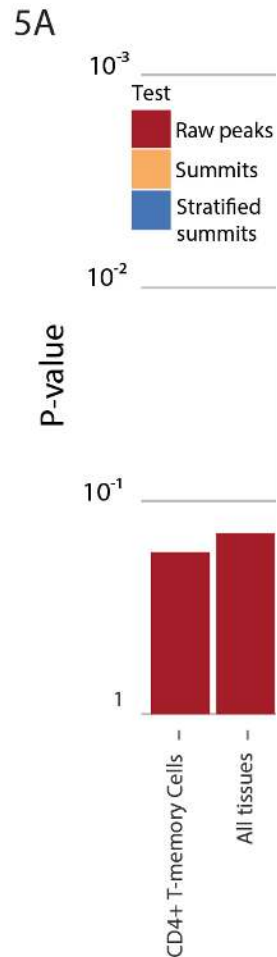
eQTLs are independently enriched for DHS and 3'UTR



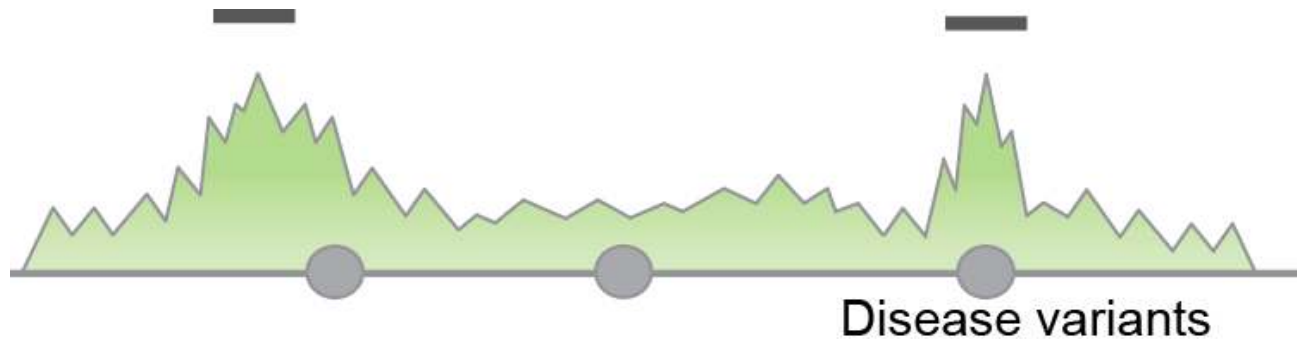
With 30 SNPs GoShifter has 80% power to detect significant enrichment



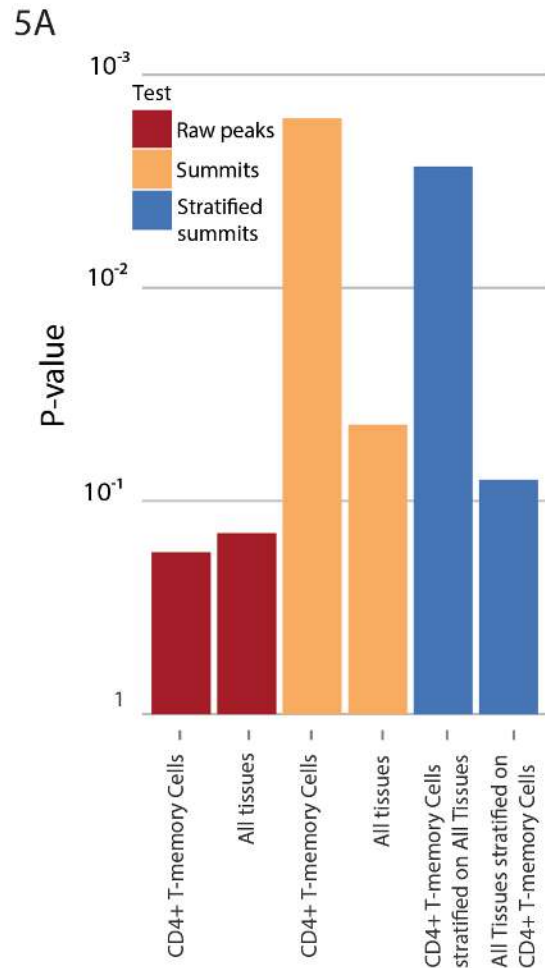
RA variants are not enriched in raw H3K4me3 peaks in CD4+ memory cells



Restricting chromatin peaks to summit regions

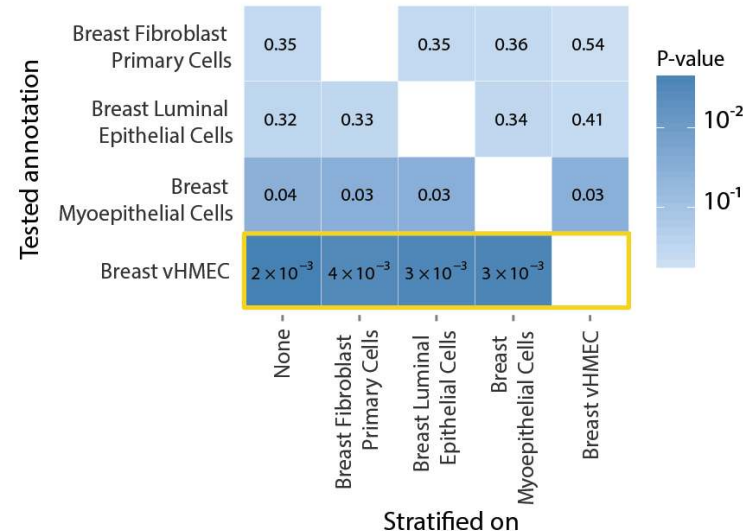
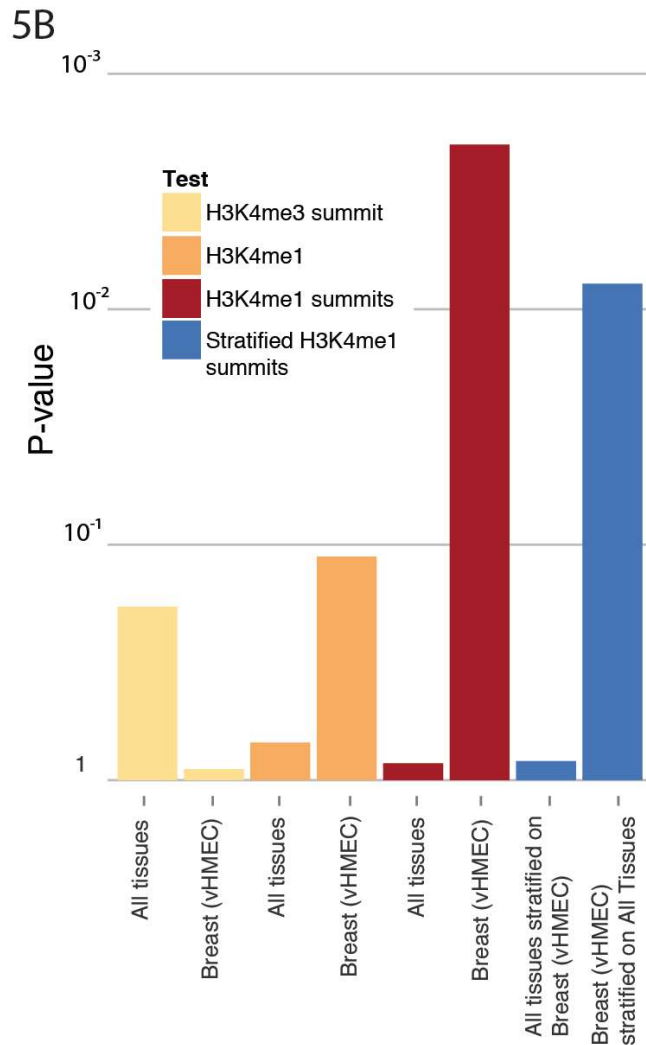


Summits of H3K4me3 peaks better capture cell type specific signatures

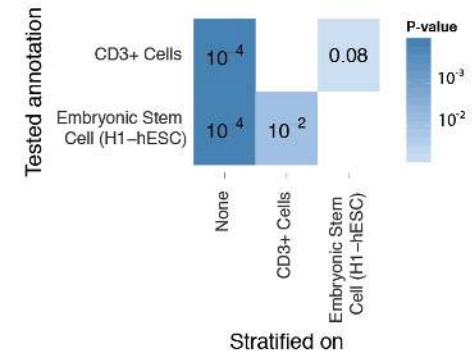
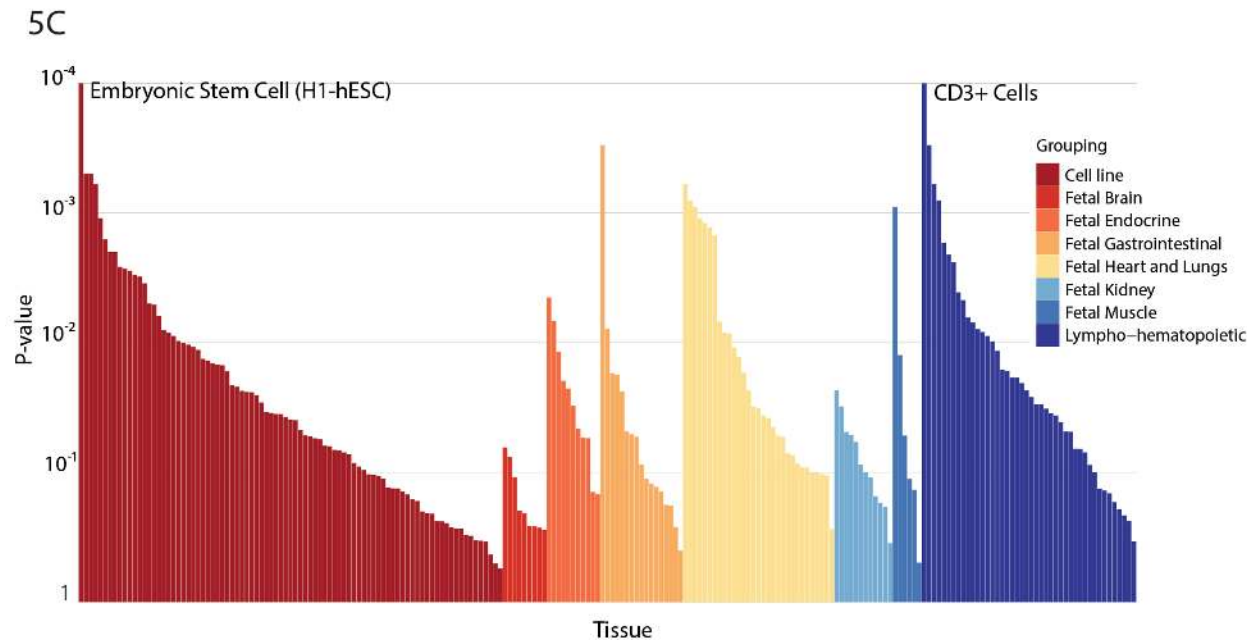


Disease enrichment at the summits of cell-type specific histone marks

Breast cancer SNPs are enriched in the summits of H3K4me1 peaks of mammary epithelial cells (vHMEC)



Stratified analysis distinguishes relevant cell types for height

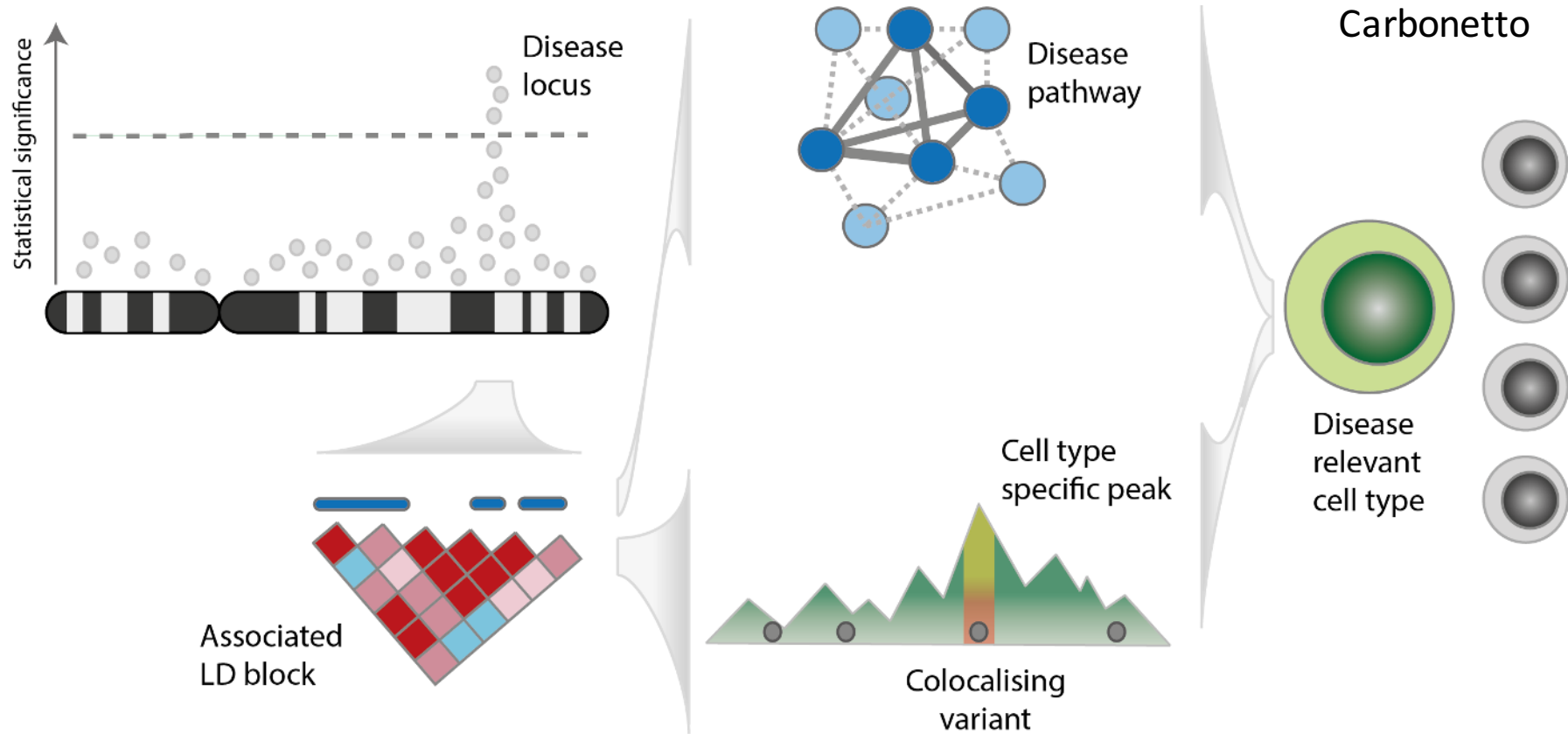


- Gene pathways represent cell type specific modules of gene expression
- If disease variants act through regulation of genes in specific pathways, stratifying enrichment within histone marks on disease pathway could improve identification of disease relevant cell types

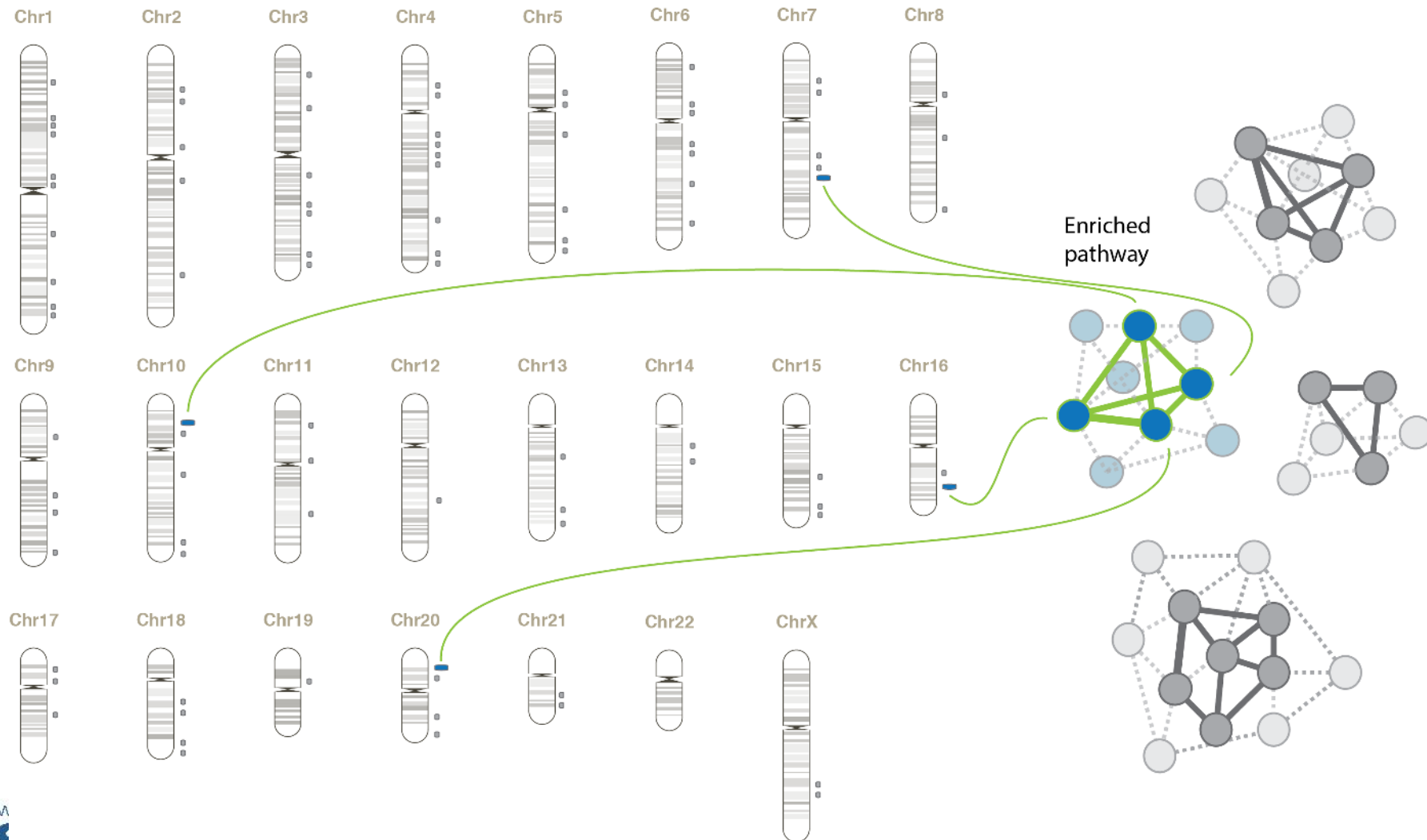
Linking disease variants to pathways and causal cell types



Peter Carbonetto



Assign SNPs to genes and test for association assuming pathway enrichment



Dataset

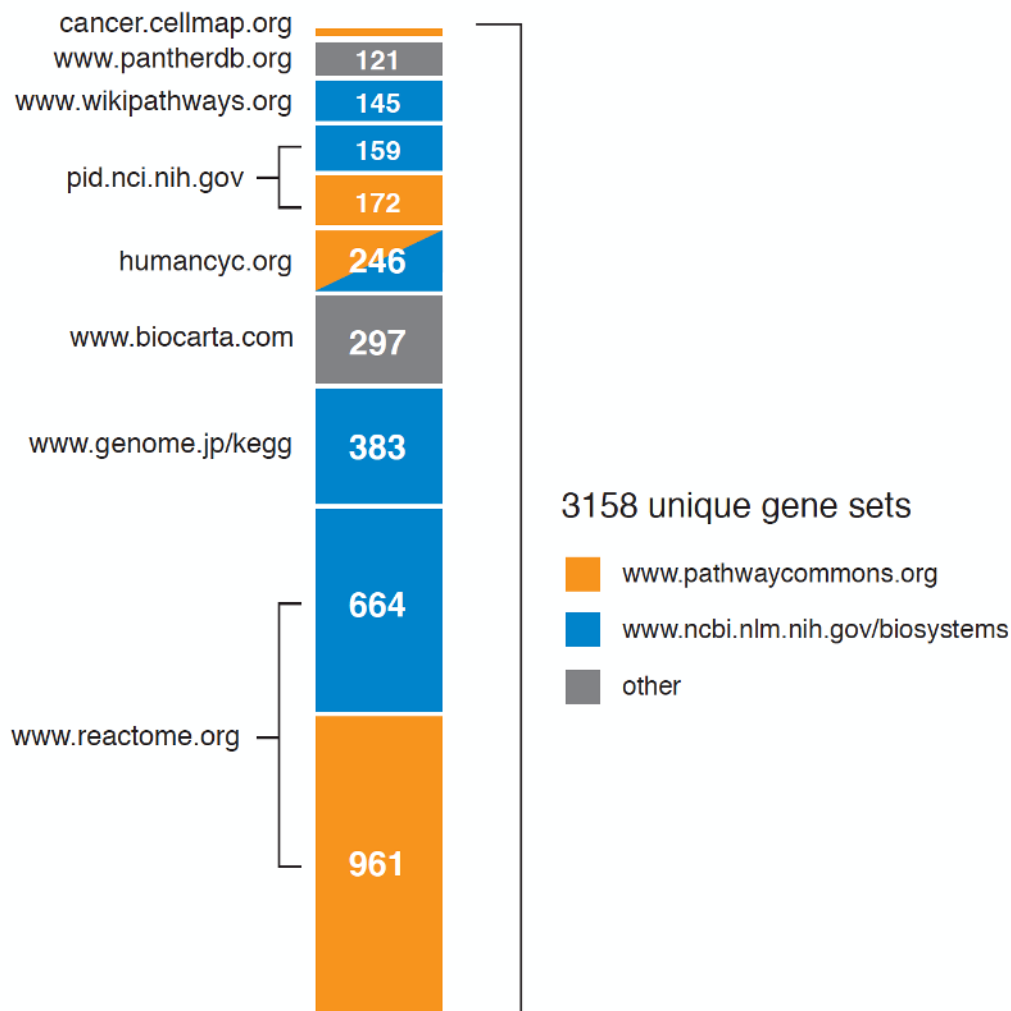
3,149 coeliac
disease cases



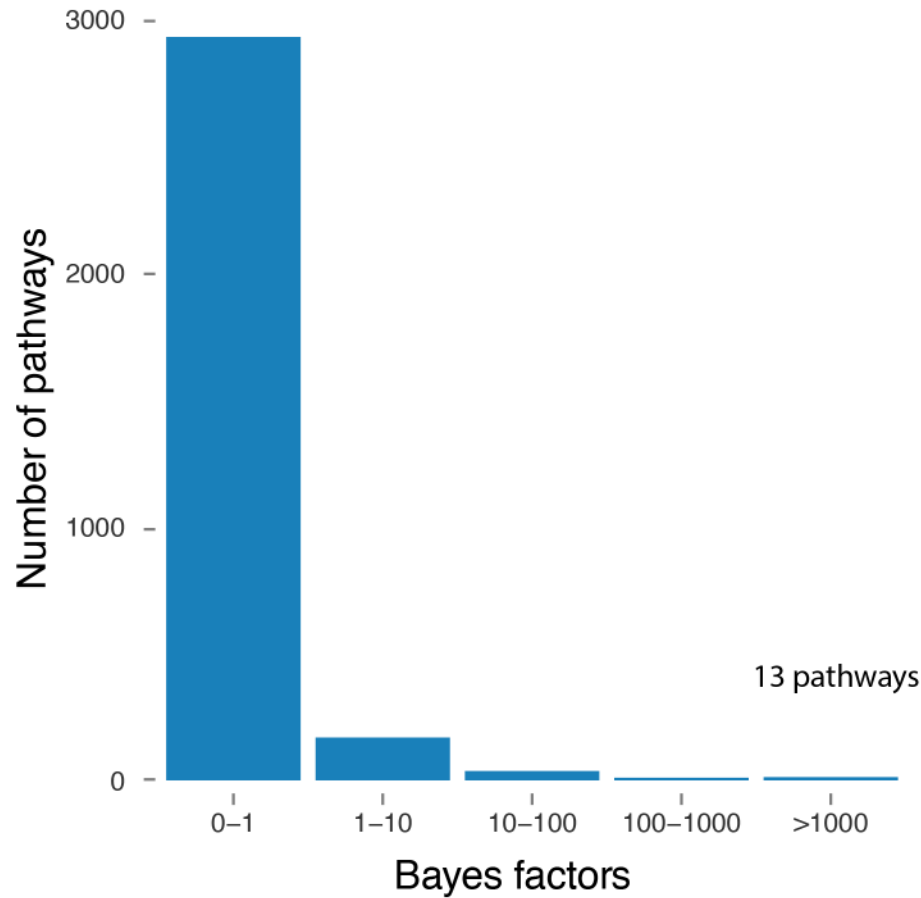
6,325 controls

Dubois et al. Nat Genet 2010

Test for enrichment 3159
candidate pathways from 8
pathway databases



Most pathways show low evidence for enrichment



Enriched pathways (Bayes factor > 10,000)

- Expression of chemokine receptors
- IL12 signaling
- Cytokine-cytokine receptor
- E-cadherin adherens junction
- Th1/Th2 differentiation

Ranking doesn't tell the full story

enriched pathway	database	source	genes	SNPs	Bayes factor
Expr of chemokine receptors	BioCarta	BioCarta	29	887	6.6e7
IL-12 signaling	PID	BioSystems	62	2,087	1.8e6
Cytokine-cytokine receptor	KEGG	BioSystems	265	8,190	1.8e5
IL-12 signaling	PID	PC	113	4,915	1.2e5
E-cadherin adherens junction	PID	BioSystems	40	2,114	4.2e4
Th1/Th2 differentiation	BioCarta	BioCarta	19	771	1.7e4

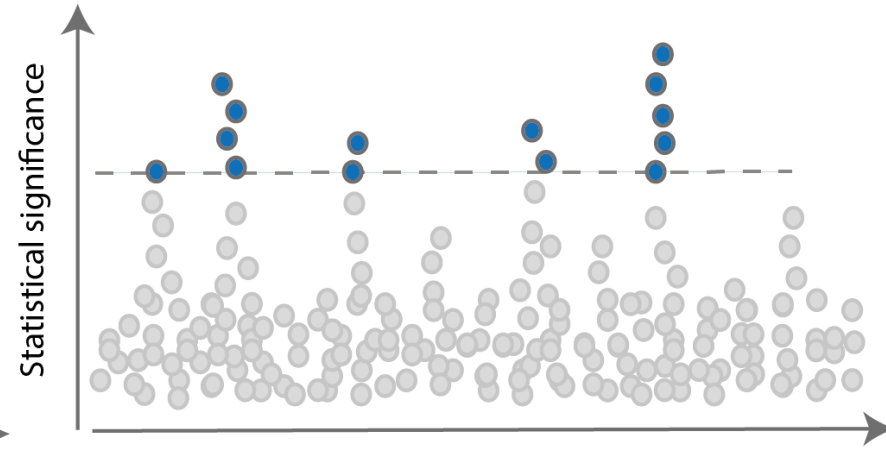
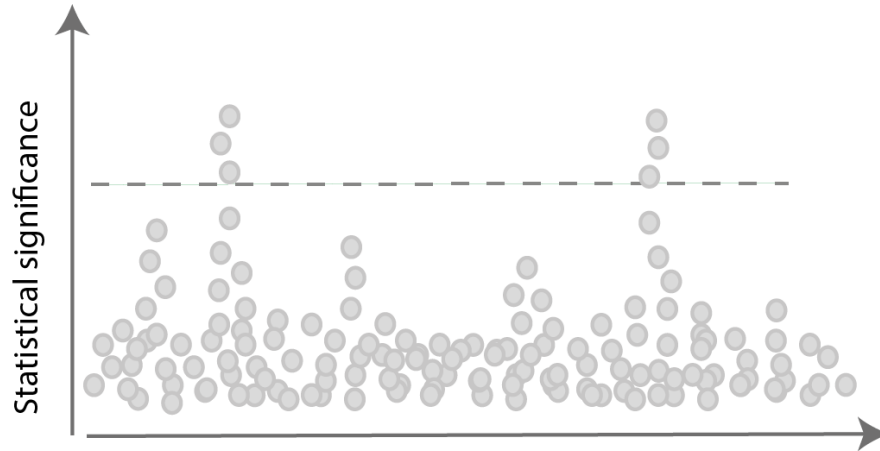
Bayes

factor enriched pathway(s)

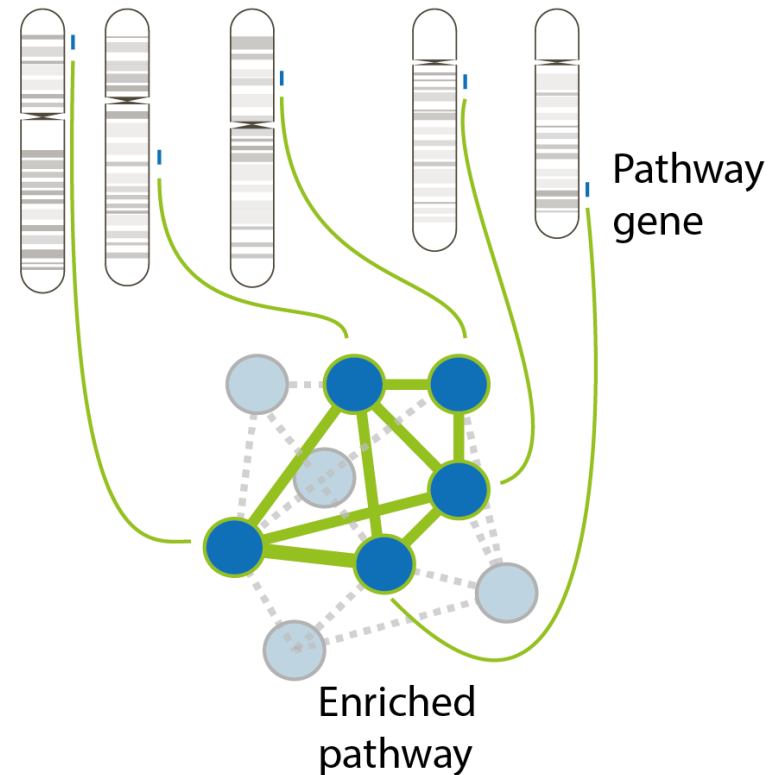
8.5e11	Expr of chemokine	+	E-cadherin adherens junction
8.7e10	IL-12 signaling	+	E-cadherin adherens junction
6.8e10	Cytokine-cytokine	+	E-cadherin adherens junction

...

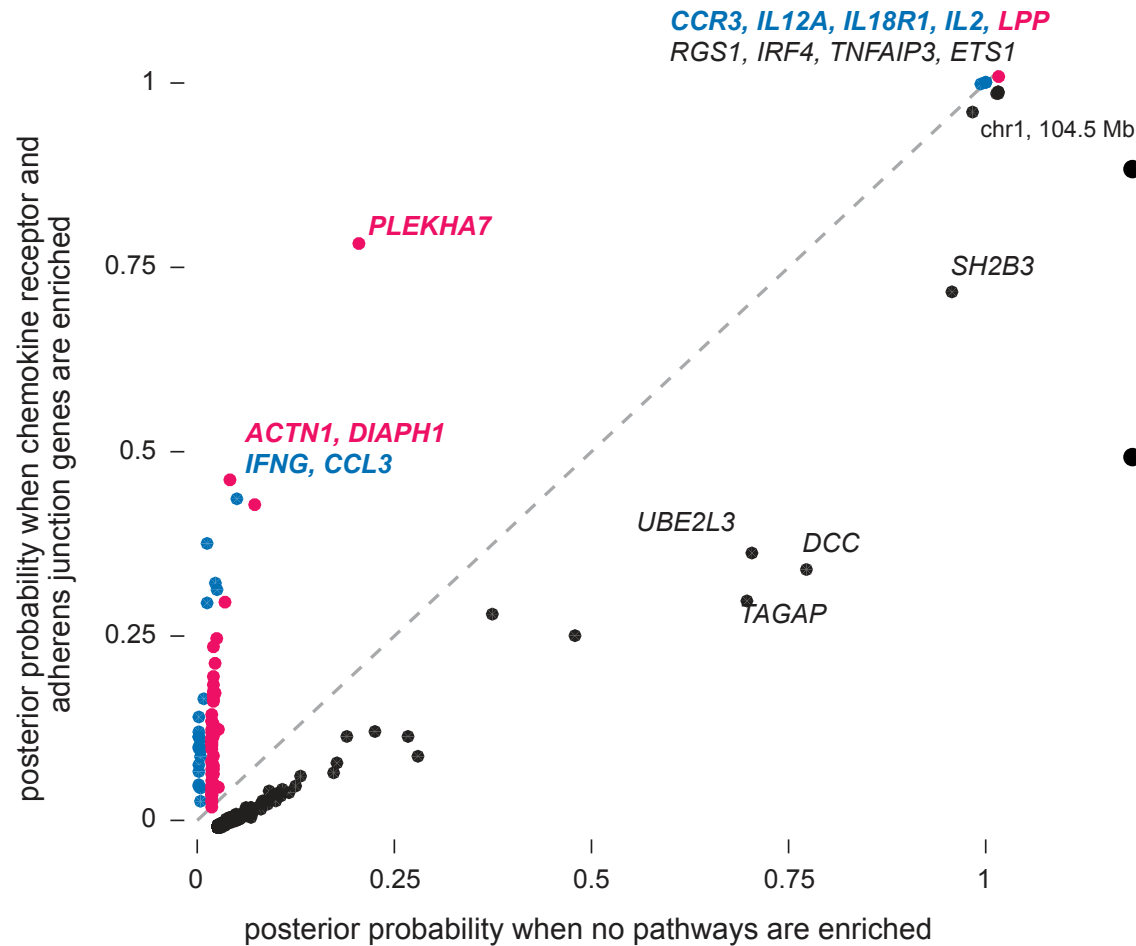
Prioritize variants within the enriched pathways



Promote new
associations assigned
to enriched pathways



Prioritisation of SNPs in enriched pathways increases support for risk factors at many loci

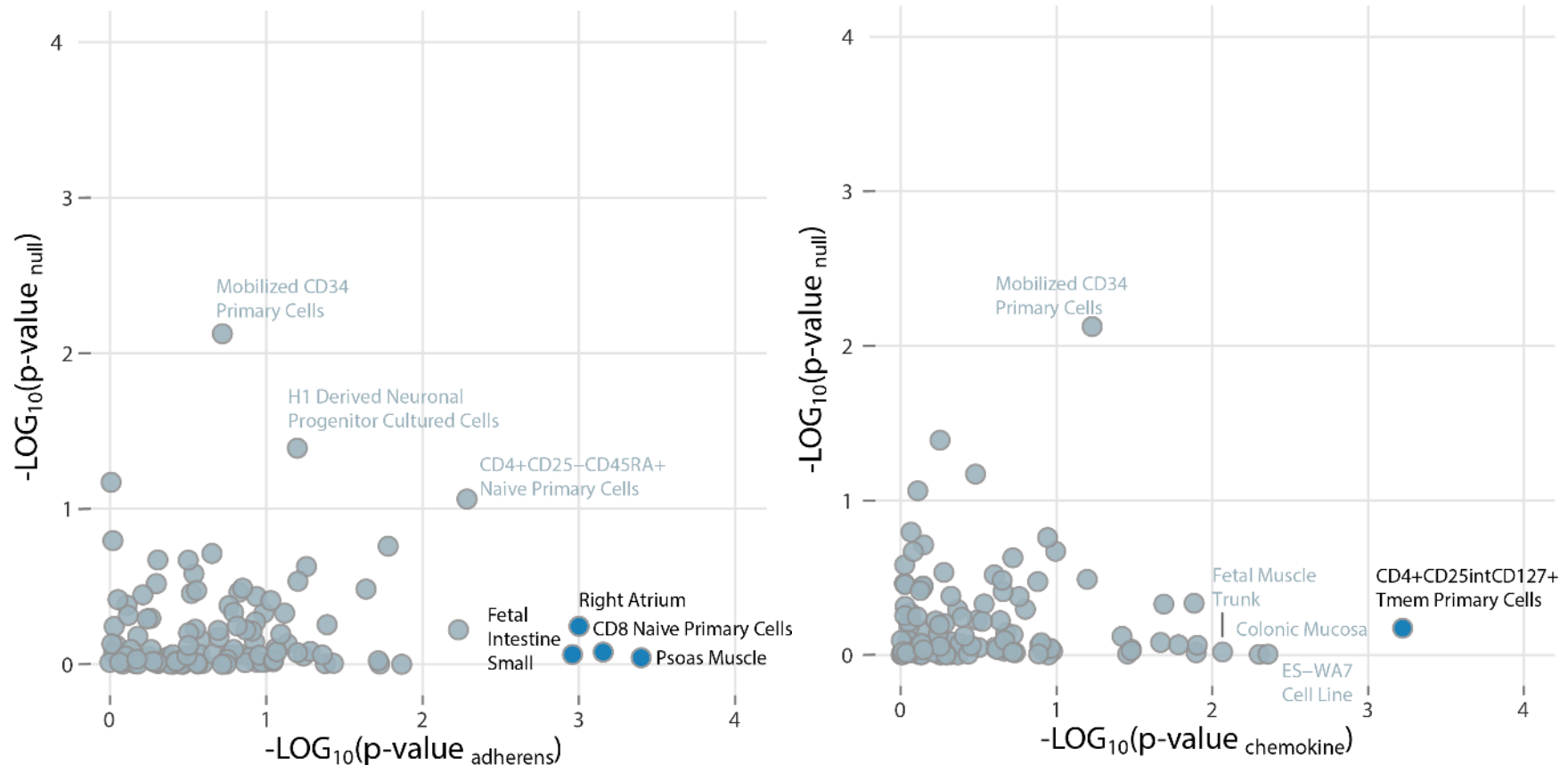


- 887 SNPs map to **chemokine receptor**
- 2,113 SNPs in **adherens junction** pathway

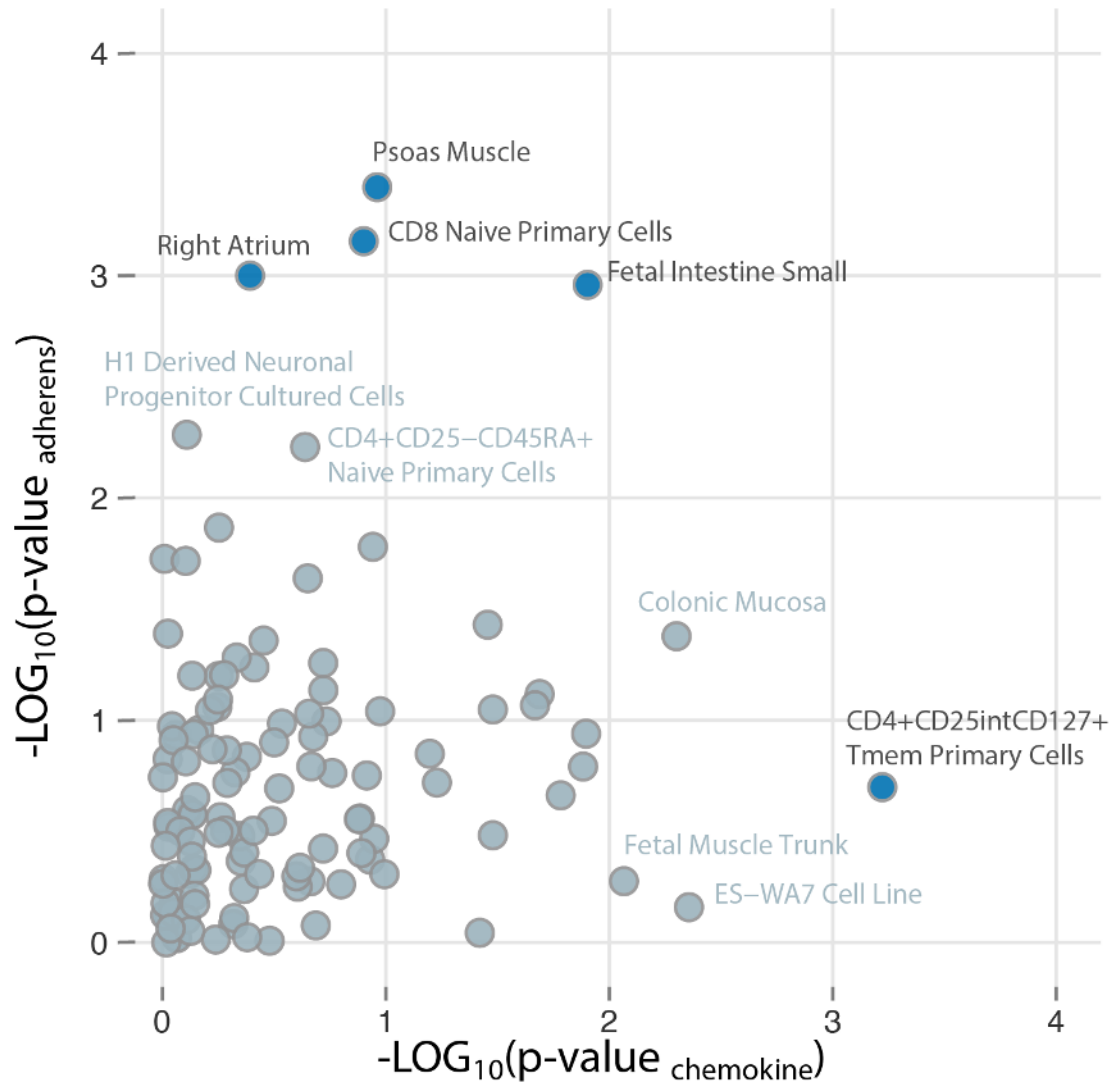
Linking pathways to effector cell types

- SNPs with posterior probability > 0.01
 - 77 SNPs assigned to chemokine receptor pathway
 - 114 SNPs assigned to adherens junction pathway
 - 158 SNPs under null of no pathway enrichment
- Total of 118 cell types assayed for H3K4me3

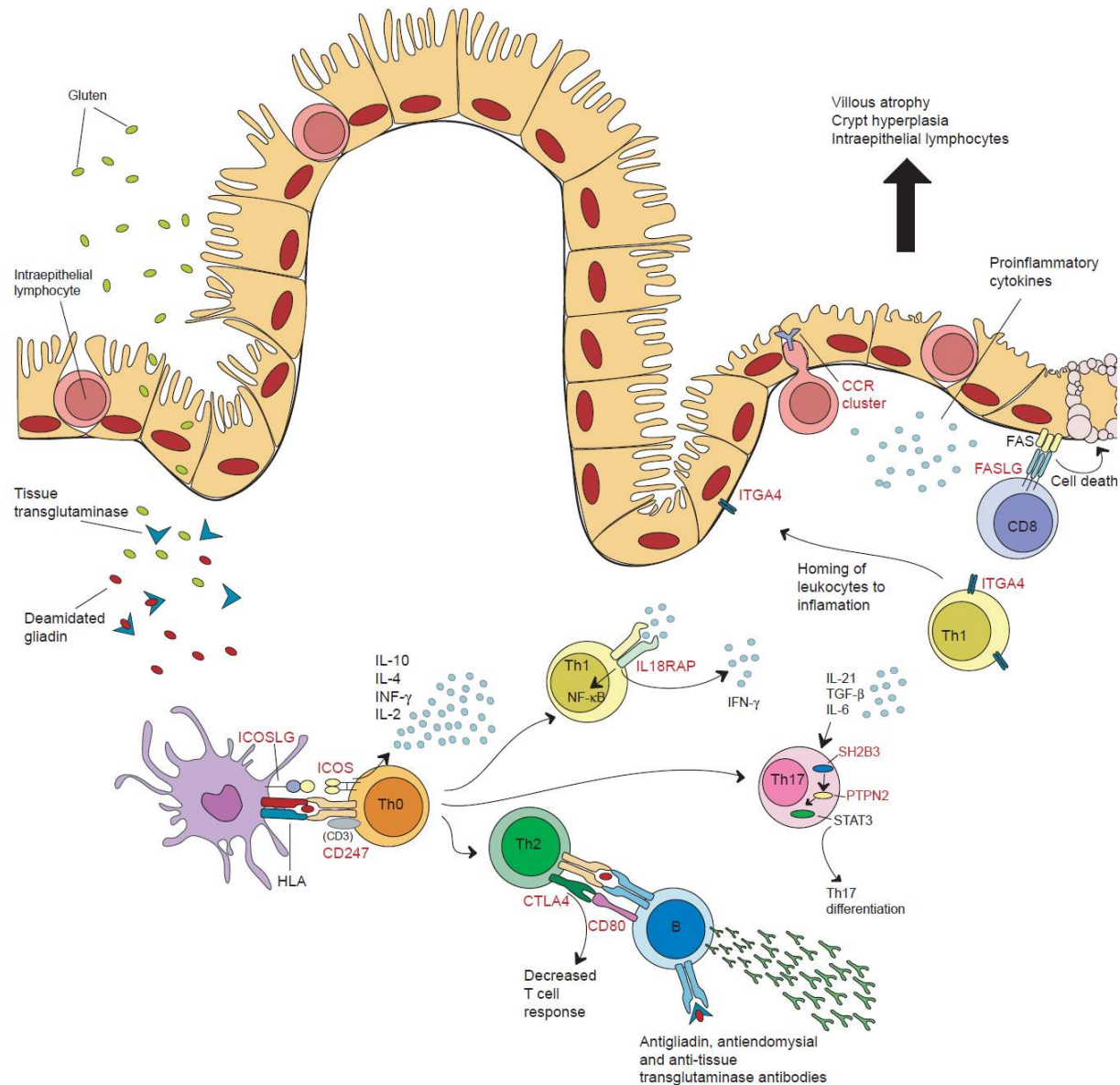
Pathway prioritise variants point to functional effects in specific cell types



Distinct pathways point to activity in different cell types



Adherens junction could point towards impaired epithelial gut barrier





Cisca Wijmenga

Groningen University, Netherlands



Peter Carbonetto

AncestryDNA, San Francisco



Soumya

Raychaudhuri

Brigham and Women's Hospital
Broad Institute



Matthew Stephens

University of Chicago

