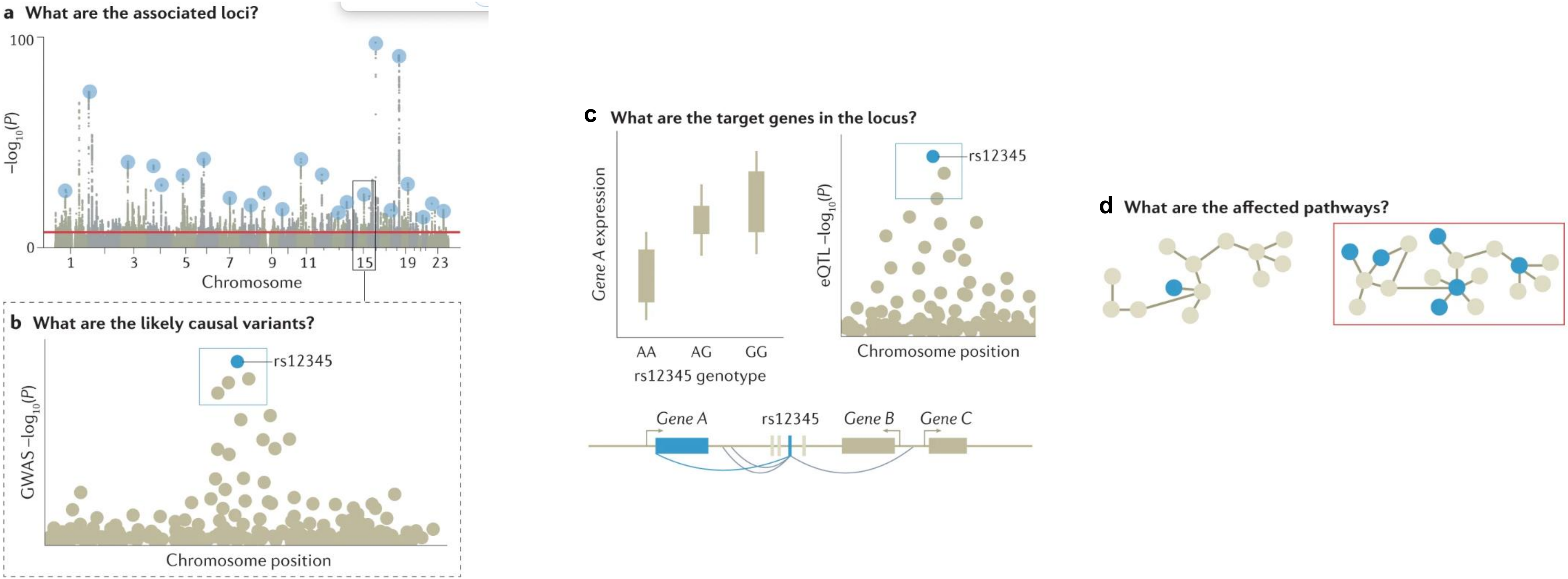
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

UQ Genetics and Genomics Winter School 2025

Systems Genomics and Pharmacogenomics
Module 6 Day 1

Functional annotation of GWAS summary data using FUMA

From GWAS signals to biological pathways, tissues and cell types



https://www.nature.com/articles/s43586-021-00056-9

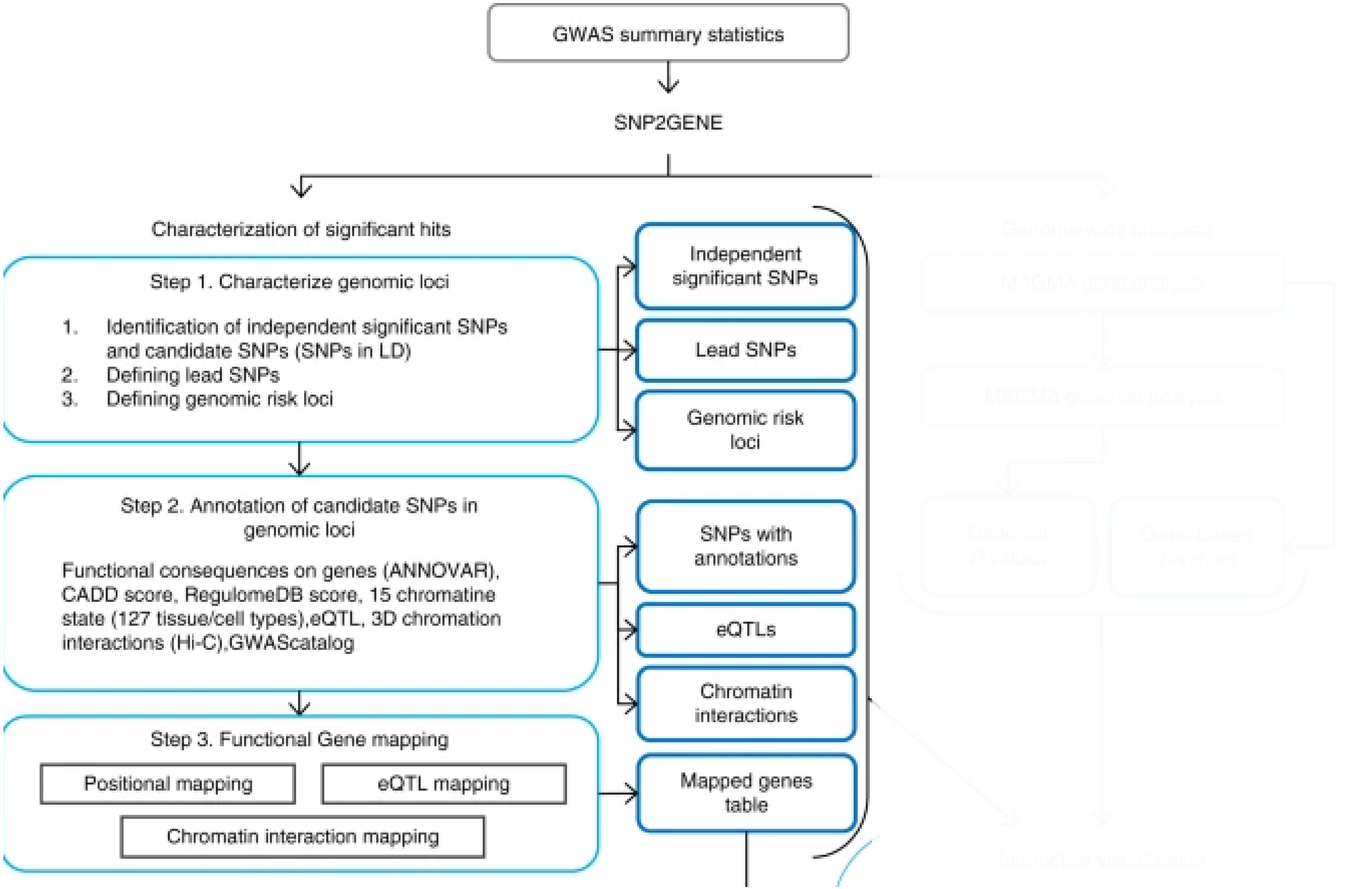
Functional mapping and annotation of genetic associations with FUMA

Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven & Danielle Posthuma ✉

Nature Communications **8**, Article number: 1826 (2017) | Cite this article

- Incorporates 18 biological repositories and tools to process GWAS summary data.

- 3 analysis modules:
  - SNP2GENE: maps GWAS SNPs to genes based on positional, eQTL and chromatin interaction
  - GENE2FUNC: biological mechanisms of prioritized genes
  - Cell type: identify cell types that may be relevant to the GWAs trait



GWAS summary statistics

SNP2GENE

Characterization of significant hits

Step 1. Characterize genomic loci

1. Identification of independent significant SNPs and candidate SNPs (SNPs in LD)
2. Defining lead SNPs
3. Defining genomic risk loci

Step 2. Annotation of candidate SNPs in genomic loci

Functional consequences on genes (ANNOVAR), CADD score, RegulomeDB score, 15 chromatine state (127 tissue/cell types),eQTL, 3D chromation interactions (Hi-C),GWAScatalog

Step 3. Functional Gene mapping

Positional mapping

eQTL mapping

Chromatin interaction mapping

Independent significant SNPs

Lead SNPs

Genomic risk loci

SNPs with annotations

eQTLs

Chromatin interactions

Mapped genes table

Genome-wide analyses

MAGMA gene analysis

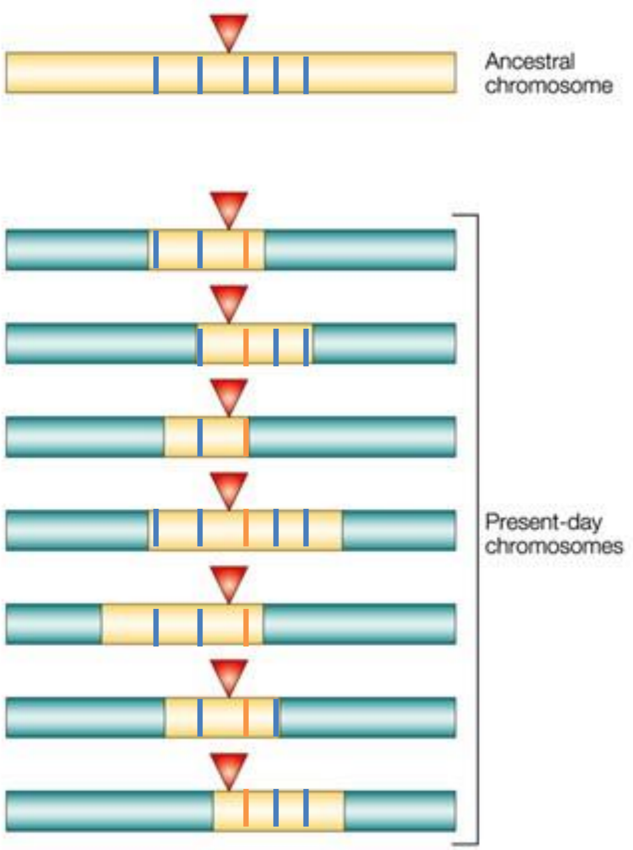
MAGMA gene-set analysis

Gene set P-values

Gene-based P-values

Interactive visualization

GWAS are based on the principle of linkage disequilibrium (LD)



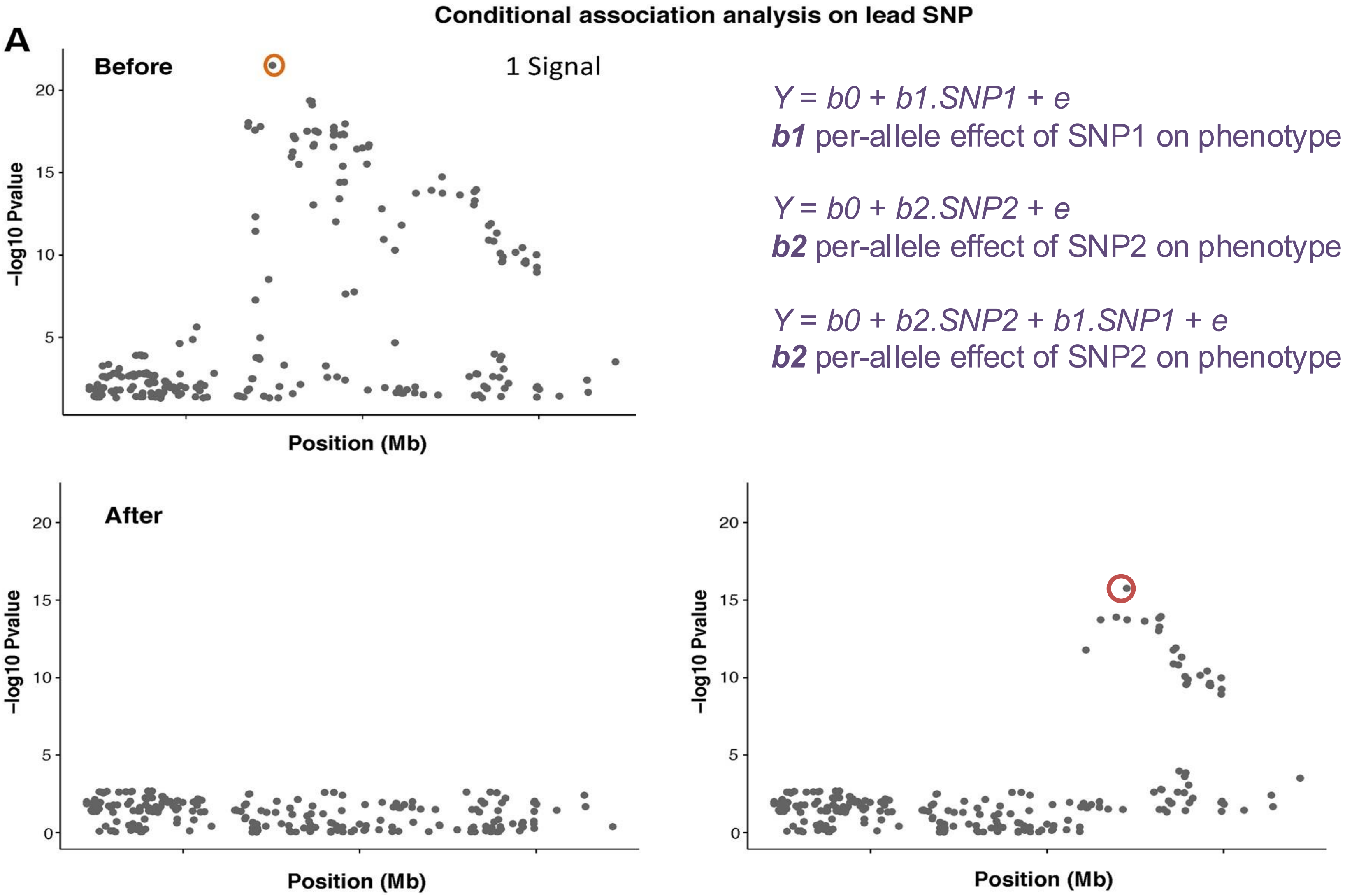
▽ Causal disease variant

| Variant in perfect LD with causal variant

Particular alleles at neighbouring regions in the genome tend to be co-inherited

A non-causal variant in LD with the causal variant will have a significant association p-value

How many risk loci are there are each GWAS association signal?



Y = b0 + b1.SNP1 + e
b1 per-allele effect of SNP1 on phenotype

Y = b0 + b2.SNP2 + e
b2 per-allele effect of SNP2 on phenotype

Y = b0 + b2.SNP2 + b1.SNP1 + e
b2 per-allele effect of SNP2 on phenotype after conditioning for SNP1

Spain & Barrett

Hum Mol Genet, 2015

Independent and candidate SNPs

1. Independent significant SNPs

- SNPs with *P*-value < 5e-8 and independent from each other at $r^2 < 0.6$ (FUMA default, can be changed)
- You can also provide your own list of SNPs to be the independent significant SNPs

2. Candidate SNPs: For each independent SNP significant, all SNPs (regardless of whether they are in input data) that have $r^2 > 0.6$ are included for further annotation. These candidate SNPs can be filtered based on user-defined MAF (MAF >=0.01 by default)

3. Independent lead SNPs: Independent significant SNPs that have $r^2 < 0.1$. If r^2 for independent sig SNPs is set to 0.1, the independent lead and independent significant SNPs will be the same.

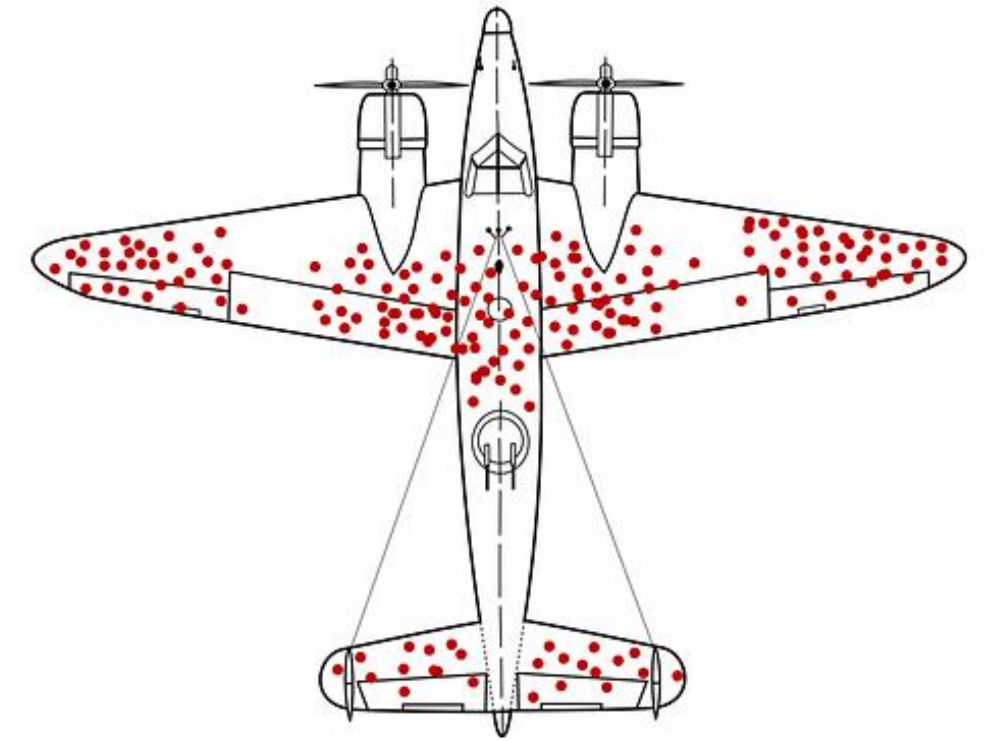
Integration of Functional Resources

Functional consequence of candidate SNPs on genes using ANNOVAR

- Combined Annotation Dependent Depletion (CADD)
- Chromatin interaction information
- The Genotype-Tissue Expression (GTEx) and other eQTL data
- Regulome DB

Combined Annotation Dependent Depletion (CADD)

A measure of variant deleteriousness (reduce organismal fitness) (Kircher et al Nature Genetics 2014) – based on the phenomenon of survivorship bias



If a mutation arises in a critical part of the genome that leads to lower survival, you are less likely to observe these in the current population.

1. Simulate all possible variants
2. Compare simulated variants with observed variants.

Deleterious variants — simulated variants that are depleted in observed data because of negative selection

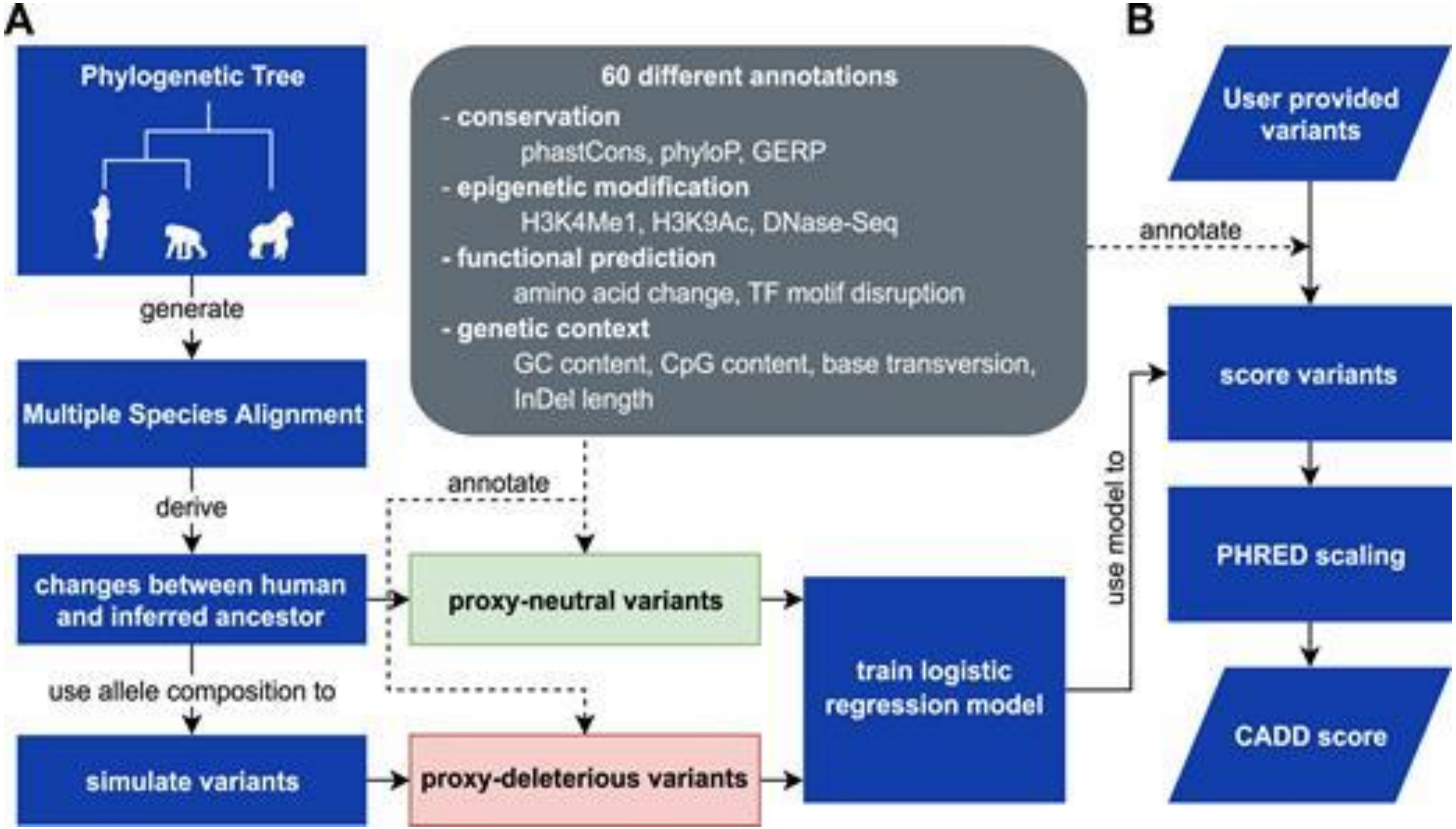
Combined Annotation Dependent Depletion (CADD)

- **Proxy-neutral variants:**
  - Variants arisen and become fixed in human populations since the split between humans and chimpanzees - mostly neutral given they have survived millions of years of purifying selection
  - Have allele frequency of 95 –100% in humans but are absent in the inferred genome sequence of the human-ape ancestor

- **Proxy-deleterious variants:**
  - Simulated *de novo* variants that would be observed in the absence of selective pressure - may include both neutral and deleterious alleles

Use these two sets of variants to identify genomic features (e.g. conservation, epigenetic modification, functional prediction) that best separates these two sets of variants

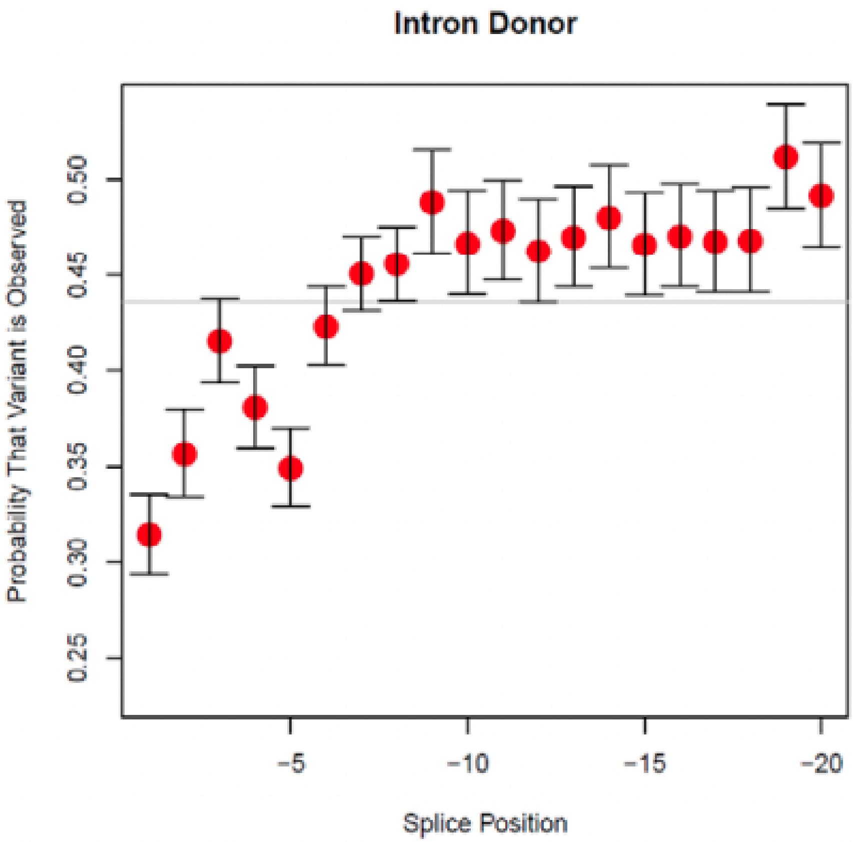
CADD


A
Phylogenetic Tree
generate
Multiple Species Alignment
derive
changes between human and inferred ancestor
use allele composition to
simulate variants
60 different annotations
- conservation
phastCons, phyloP, GERP
- epigenetic modification
H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
amino acid change, TF motif disruption
- genetic context
GC content, CpG content, base transversion, InDel length
annotate
proxy-neutral variants
proxy-deleterious variants
train logistic regression model
use model to
B
User provided variants
annotate
score variants
PHRED scaling
CADD score

CADD

Genomic features predictive of deleteriousness:

- ~20-fold depletion of **nonsense variants**
- ~2-fold depletion of **missense variants**
- no depletion of intergenic or upstream or downstream variants
- Nonsense and missense mutations that occurred near the start sites of coding DNA were more depleted than those occurring near the ends
- Variants within 20, and especially within 2, nucleotides of splice junctions were also depleted



A scaled score of 10 or greater indicates a raw score in the top 10% of all possible reference genome SNVs, regardless of the details of the annotation

A score of 20 or greater indicates a raw score in the top 1% of all possible reference genome SNVs, regardless of the details of the annotation

eQTL mapping – mostly cis-regulation

- GTEx
- EyeGEx (retina in 406 individuals)
- eQTL catalogue
- eQTLGen (~31,000 samples European) http://www.eqtlgen.org/index.html
- Blood eQTL Westra et al 2013 (~5300 blood samples from 7 studies)
- PsychENCODE (brain data ~1400 samples) http://resource.psychencode.org
- BIOS QTL browser (~2000 whole blood healthy adults from 4 Dutch cohorts Zhernakova et al. 2017)
- Braineac (Brain expression in 134 controls of European ancestry) http://www.braineac.org/

Chromatin interaction

- Identifying regions of DNA that physically interact with each other
- Interaction between distal regulatory elements with promoters to regulate gene expression

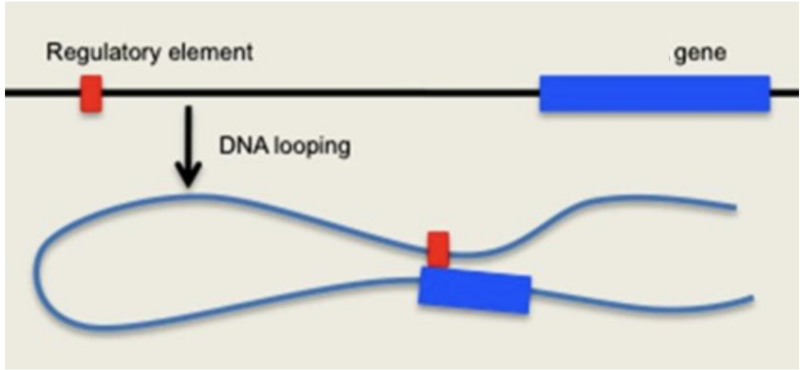


Figure DOI: 10.3389/fnmol.2013.00032

Cell Rep. Author manuscript; available in PMC 2017 Jun 20.

Published in final edited form as:

Cell Rep. 2016 Nov 15; 17(8): 2042–2059.

doi: 10.1016/j.celrep.2016.10.061

PMCID: PMC5478386

NIHMSID: NIHMS828671

PMID: 27851967

A Compendium of Chromatin Contact Maps Reveal Spatially Active Regions in the Human Genome

Anthony D. Schmitt,[1,2,10,*] Ming Hu,[3,11,*#] Inkyung Jung,[1,12] Zheng Xu,[4,13] Yunjiang Qiu,[1,5] Catherine L. Tan,[1,10] Yun Li,[4] Shin Lin,[6] Yiing Lin,[7] Cathy L. Barr,[8] and Bing Ren[1,9,#]

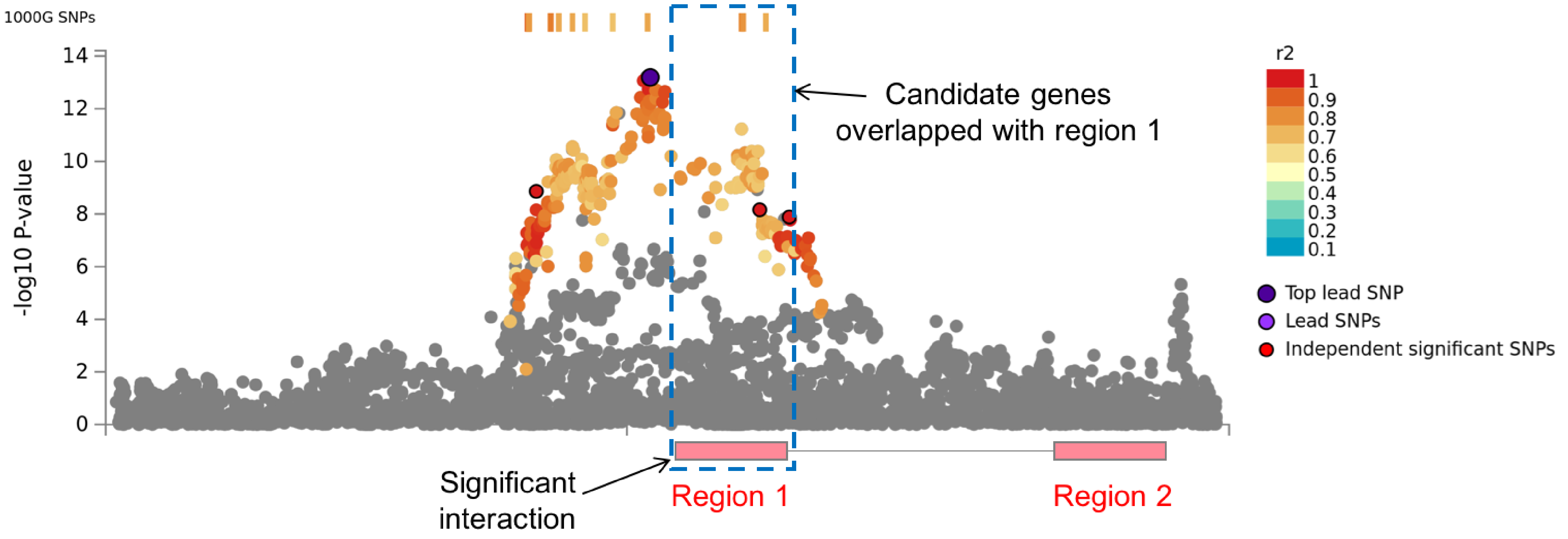
▸ Author information ▸ Copyright and License information Disclaimer

- Chromatin interactions maps in 21 primary human tissues and cell types
- Identified genomic regions that exhibit unusually high levels of interaction (frequently interacting regions or FIRE)
- FIREs enriched for super-enhancers and are near cell-identity genes
- FIREs conserved in human and mouse
- FIREs enriched for disease-associated GWAS SNPs

Chromatin interaction

Region 1: One end of the interaction that overlaps with one of the candidate SNPs

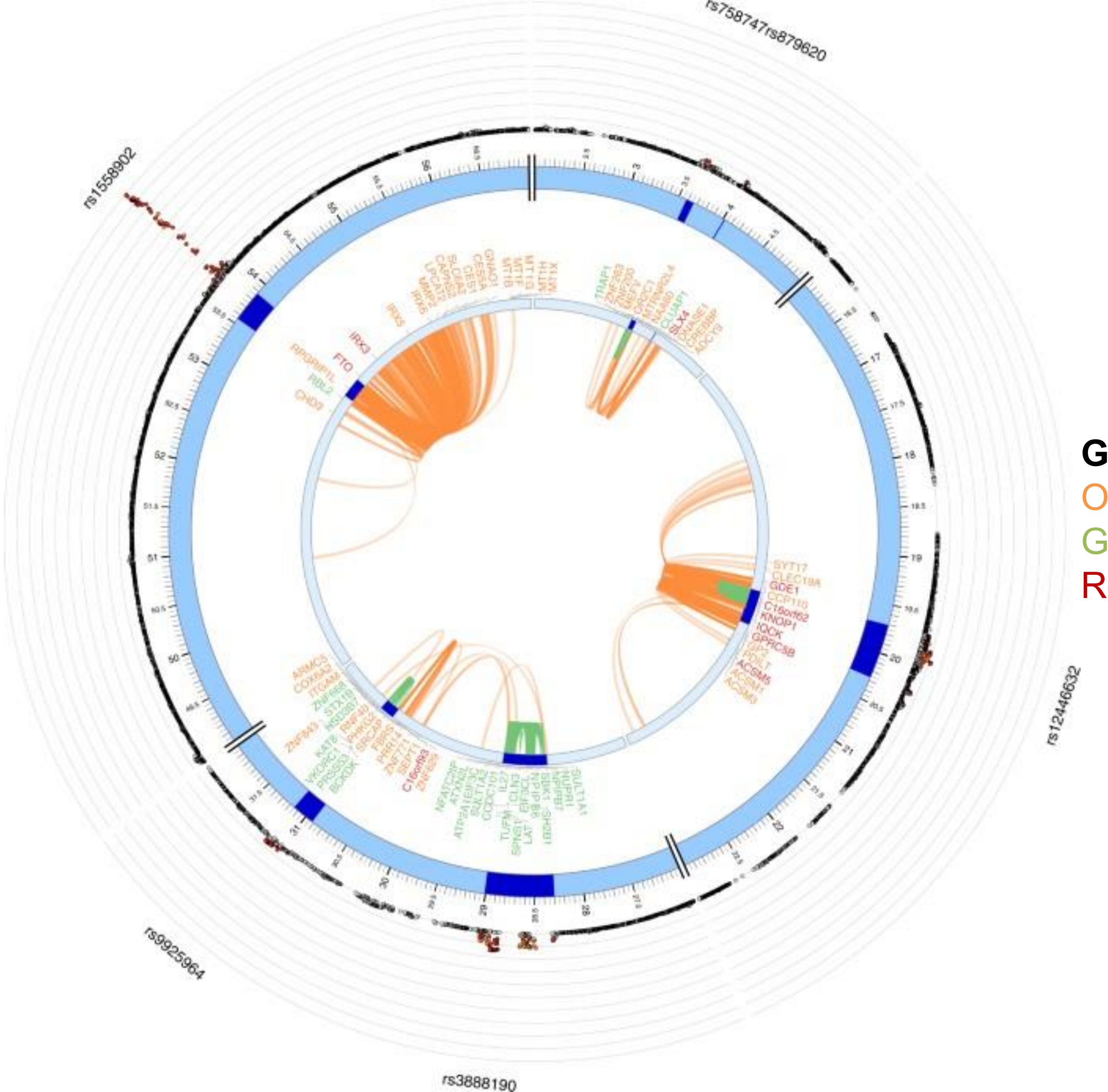
Region 2: Other end of the significant interaction. Identifies genes whose promoter region interacts with the region containing the candidate SNPs



Candidate genes overlapped with region 1

Significant interaction

Chromatin interactions and eQTLS of a BMI risk locus on chr16



Genes

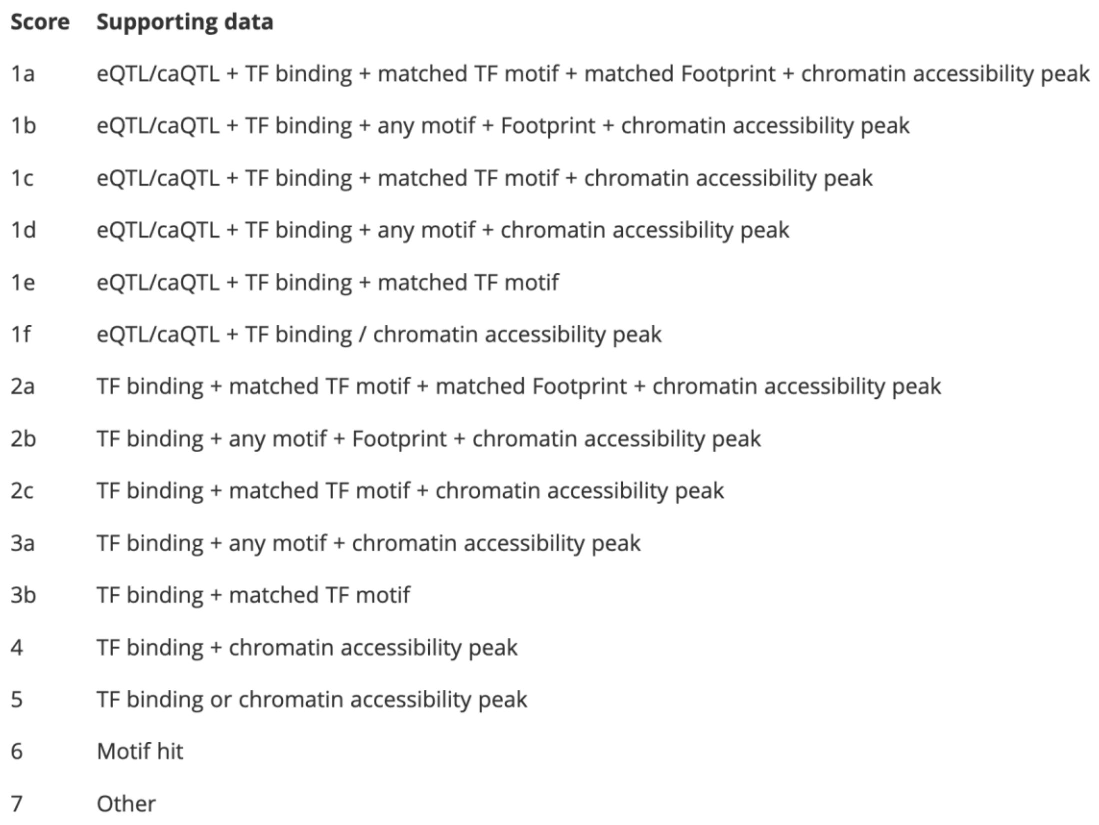
Orange: mapped by eQTL data

Green: mapped by HiC data

Red: mapped by both

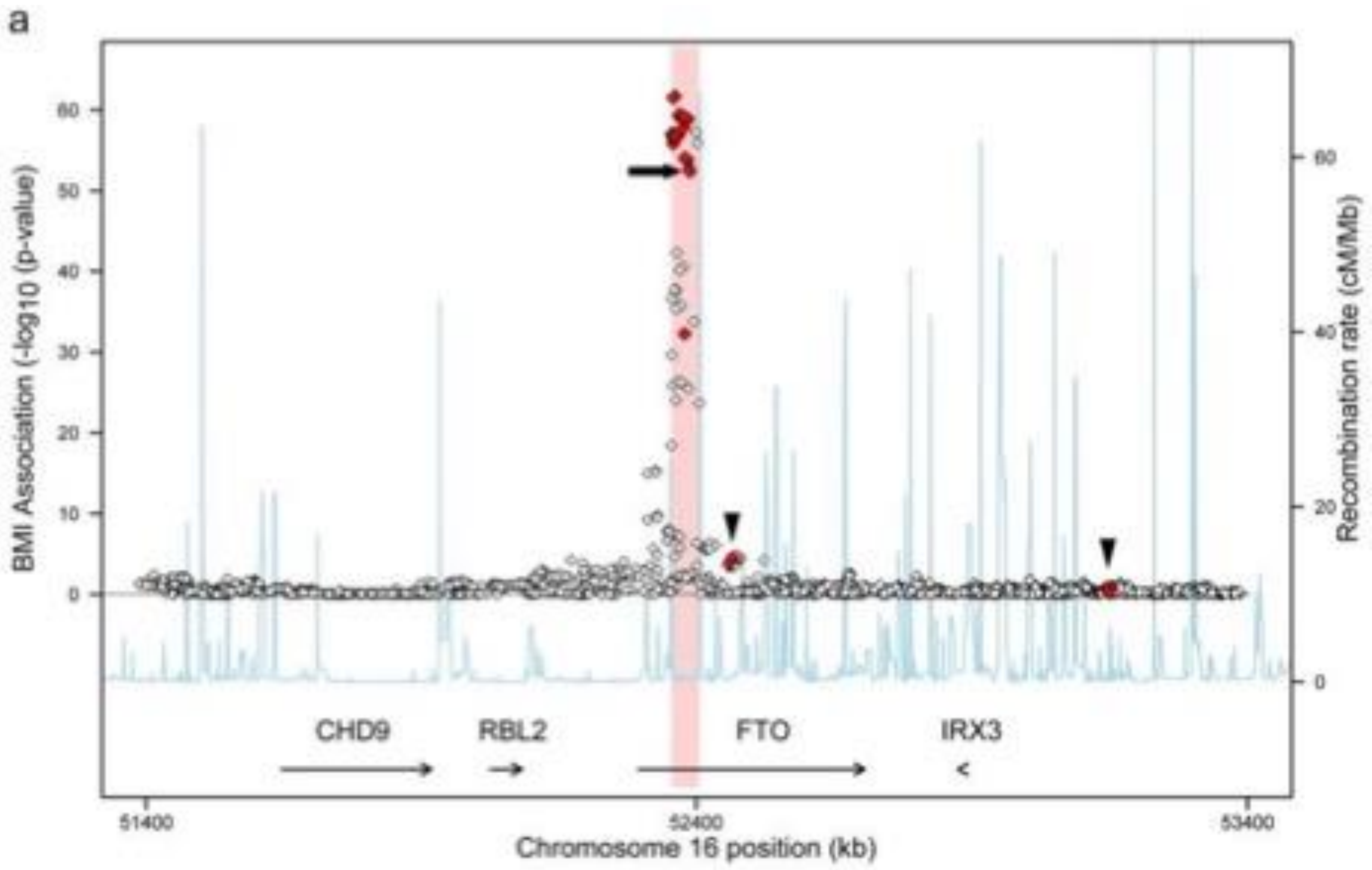
RegulomeDB

- Intersects candidate SNPs with known functionally-active regions identified from functional genomic assays e.g. TF ChiP-seq (TF-binding regions), DHS (open chromatin regions)
- Scores functional consequence of each SNP based on strength of evidence



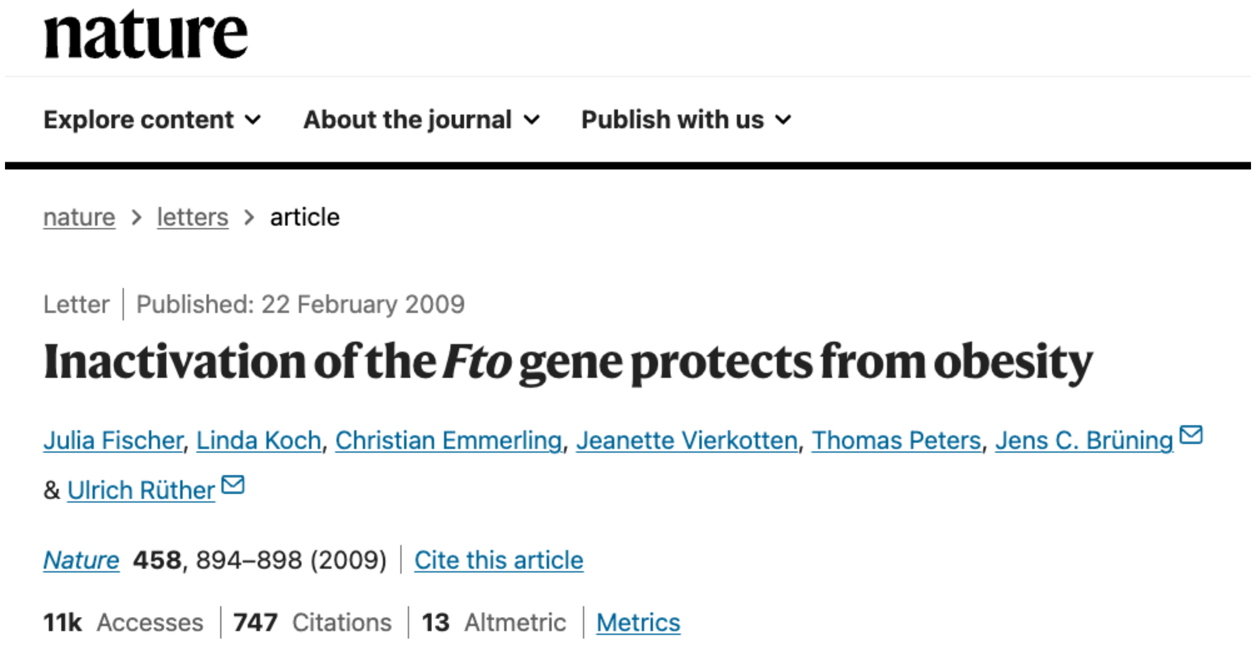
Score	Supporting data
1a	eQTL/caQTL + TF binding + matched TF motif + matched Footprint + chromatin accessibility peak
1b	eQTL/caQTL + TF binding + any motif + Footprint + chromatin accessibility peak
1c	eQTL/caQTL + TF binding + matched TF motif + chromatin accessibility peak
1d	eQTL/caQTL + TF binding + any motif + chromatin accessibility peak
1e	eQTL/caQTL + TF binding + matched TF motif
1f	eQTL/caQTL + TF binding / chromatin accessibility peak
2a	TF binding + matched TF motif + matched Footprint + chromatin accessibility peak
2b	TF binding + any motif + Footprint + chromatin accessibility peak
2c	TF binding + matched TF motif + chromatin accessibility peak
3a	TF binding + any motif + chromatin accessibility peak
3b	TF binding + matched TF motif
4	TF binding + chromatin accessibility peak
5	TF binding or chromatin accessibility peak
6	Motif hit
7	Other

GWAS to mechanism – the *FTO* story



- The FTO locus - first ever GWAS locus to be associated with obesity in 2007
- Individuals homozygous for the top risk variant weigh ~3kg more than non-carriers.

GWAS to mechanism – the *FTO* story



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › letters › article

Letter | Published: 22 February 2009

Inactivation of the *Fto* gene protects from obesity

Julia Fischer, Linda Koch, Christian Emmerling, Jeanette Vierkotten, Thomas Peters, Jens C. Brüning ✉ & Ulrich Rüther ✉

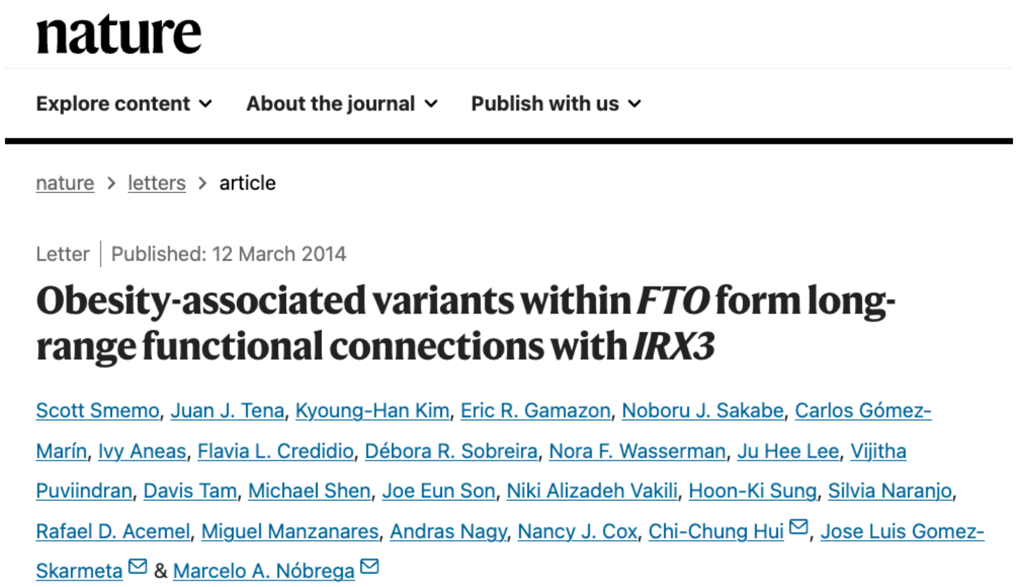
Nature **458**, 894–898 (2009) | Cite this article

11k Accesses | **747** Citations | **13** Altmetric | Metrics

FTO gene was the primary suspect

Fto knockout mice were stunted and lean, and the leanness was mainly due to burning too much fat.

GWAS to mechanism – the *FTO* story



Letter | Published: 12 March 2014

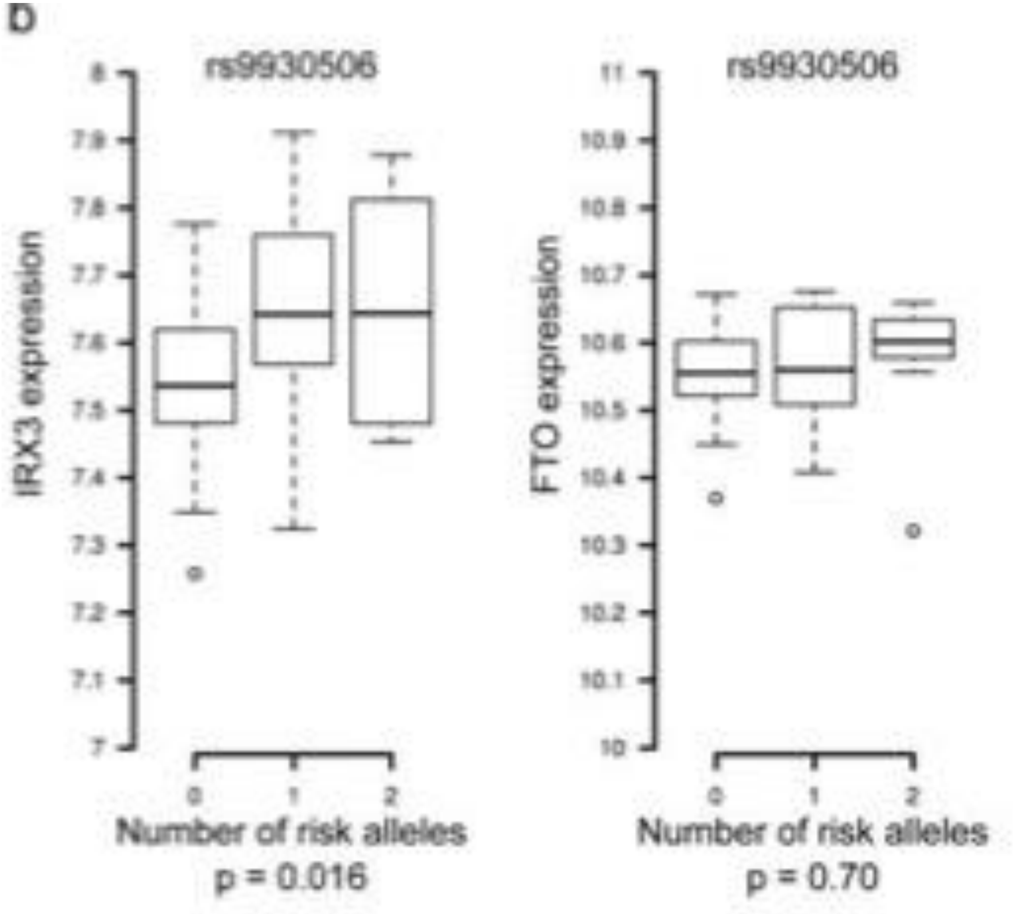
Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*

Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L. Credidio, Débora R. Sobreira, Nora F. Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D. Acemel, Miguel Manzanares, Andras Nagy, Nancy J. Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta & Marcelo A. Nóbrega

Chromatin interaction data

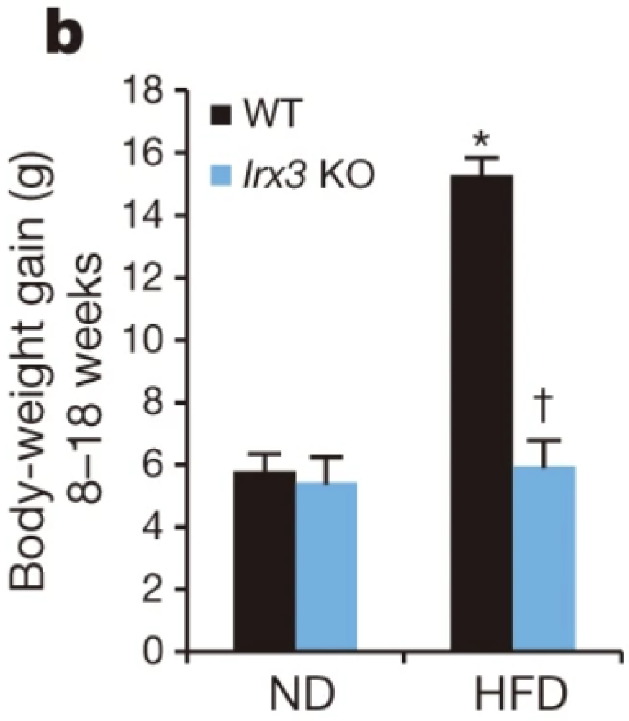
Promoter of *Irx3* participates in numerous long-range interactions, including with the GWAS region in both mouse embryo and adult mouse brain, as well as MCF-7 cells and zebrafish embryos

GWAS to mechanism – the *FTO* story

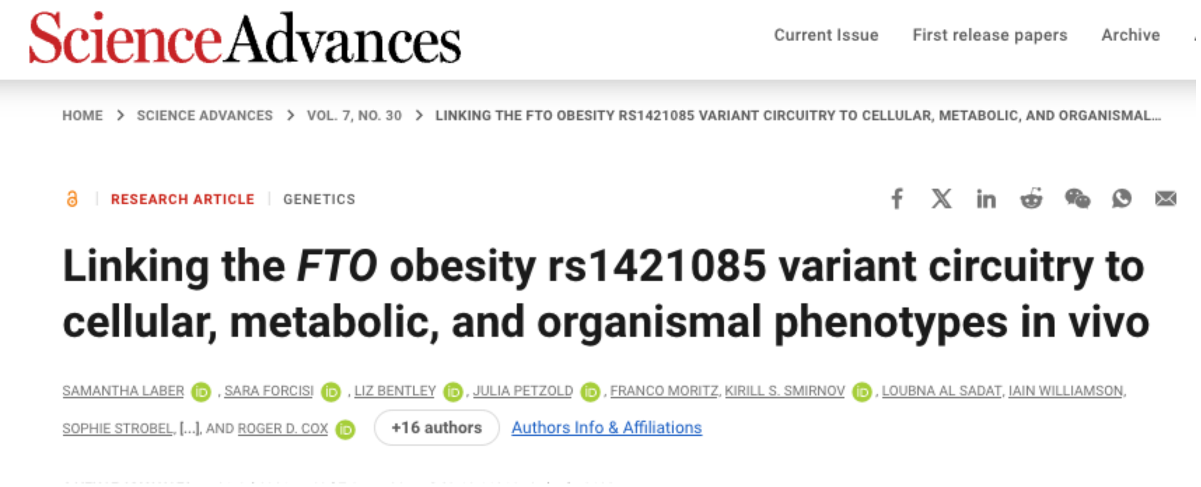


BMI-associated SNPs are eSNPs for *IRX3*, not *FTO*, expression in human brain

***Irx3*-deficient mice are leaner and are protected against diet-induced obesity**



GWAS to mechanism – the *FTO* story



HOME > SCIENCE ADVANCES > VOL. 7, NO. 30 > LINKING THE FTO OBESITY RS1421085 VARIANT CIRCUITRY TO CELLULAR, METABOLIC, AND ORGANISMAL...

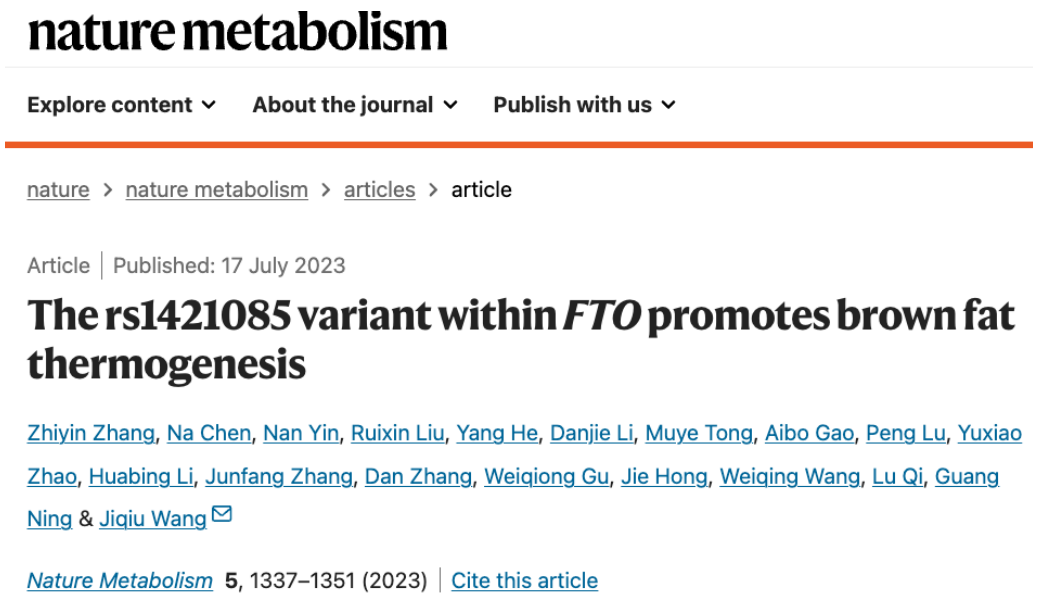
RESEARCH ARTICLE | GENETICS

Linking the *FTO* obesity rs1421085 variant circuitry to cellular, metabolic, and organismal phenotypes in vivo

SAMANTHA LABER, SARA FORCISI, LIZ BENTLEY, JULIA PETZOLD, FRANCO MORITZ, KIRILL S. SMIRNOV, LOUBNA AL SADAT, IAIN WILLIAMSON, SOPHIE STROBEL, [...], AND ROGER D. COX +16 authors Authors Info & Affiliations

Instead of knocking out genes, deleted the non-coding region in the FTO gene
Mice don't gain weight when fed with a high-fat diet, and deleting this locus increases *Irx3* and *Irx4* expression.

GWAS to mechanism – the *FTO* story



Article | Published: 17 July 2023

The rs1421085 variant within *FTO* promotes brown fat thermogenesis

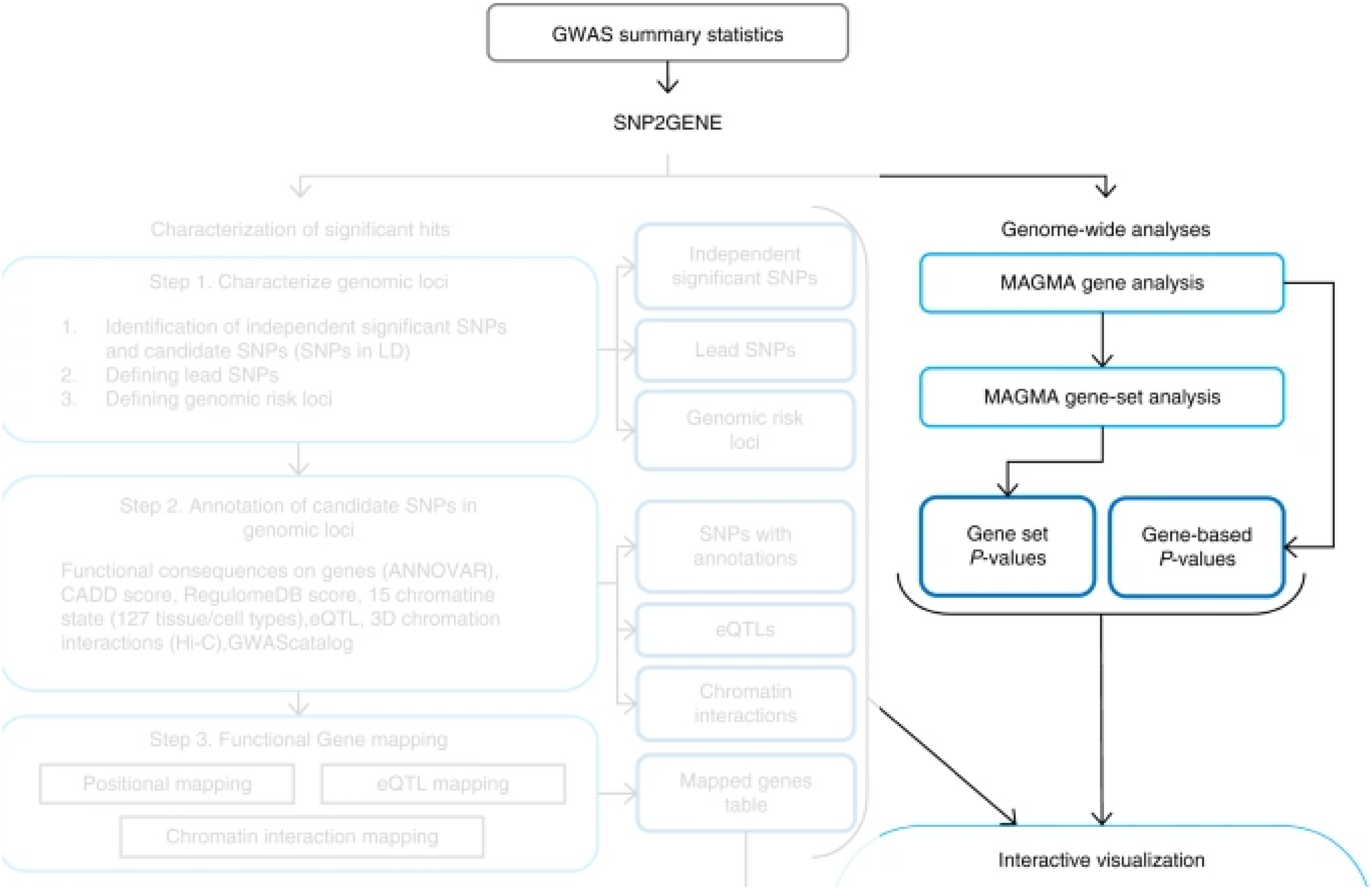
Zhiyin Zhang, Na Chen, Nan Yin, Ruixin Liu, Yang He, Danjie Li, Muye Tong, Aibo Gao, Peng Lu, Yuxiao Zhao, Huabing Li, Junfang Zhang, Dan Zhang, Weiqiong Gu, Jie Hong, Weiqing Wang, Lu Qi, Guang Ning & Jiqiu Wang

Nature Metabolism **5**, 1337–1351 (2023) | Cite this article

- Recreate the exact genetic variant in mice and study the consequences.
- The risk allele, that increased weight in humans, decreased weight in mice.
- Effect of variant is temperature-dependent:
  - At room temperature (22°C, which is ambient for humans but not mice) mice were resistant to high-fat diet (HFD) induced obesity.
  - At 29–31°C (ambient for mice), the effects of the variant were ameliorated.
- rs1421085 T>C has a role in improving survival in cold conditions, as it enhances brown adipose thermogenesis.

Lessons from the *FTO* story

- Extrapolation of findings in animal models to humans
- GWAS to mechanism is a long and winding road!
- Biology is extremely complicated!
- Context-dependent variant effects



GWAS summary statistics

SNP2GENE

Characterization of significant hits

Step 1. Characterize genomic loci

1. Identification of independent significant SNPs and candidate SNPs (SNPs in LD)
2. Defining lead SNPs
3. Defining genomic risk loci

Independent significant SNPs

Lead SNPs

Genomic risk loci

Step 2. Annotation of candidate SNPs in genomic loci

Functional consequences on genes (ANNOVAR), CADD score, RegulomeDB score, 15 chromatine state (127 tissue/cell types),eQTL, 3D chromation interactions (Hi-C),GWAScatalog

SNPs with annotations

eQTLs

Chromatin interactions

Step 3. Functional Gene mapping

Positional mapping

eQTL mapping

Chromatin interaction mapping

Mapped genes table

Genome-wide analyses

MAGMA gene analysis

MAGMA gene-set analysis

Gene set *P*-values

Gene-based *P*-values

Interactive visualization

Gene-based tests

- GWAS focus on a single genetic variant with a trait at a time
  - Large multiple-testing burden

- Gene-based tests - testing joint association of all markers in a gene with the phenotype
  - Reduced multiple-testing burden (millions of SNPs vs ~22,000 genes)
  - Detect effects consisting of multiple weaker associations

- Several methods available – PLINK, **MAGMA** (implemented in FUMA), fastBAT
  - Simplest approach – combine p-values or $\chi 2$-statistics estimated for each variant within the region of interest
  - Need to account for SNP correlation structure

    - Summary-based tests require a reference dataset (of similar ancestry) for estimating SNP-SNP correlations

Gene-based association test - MAGMA

Step 1: Mapping SNPs to gene

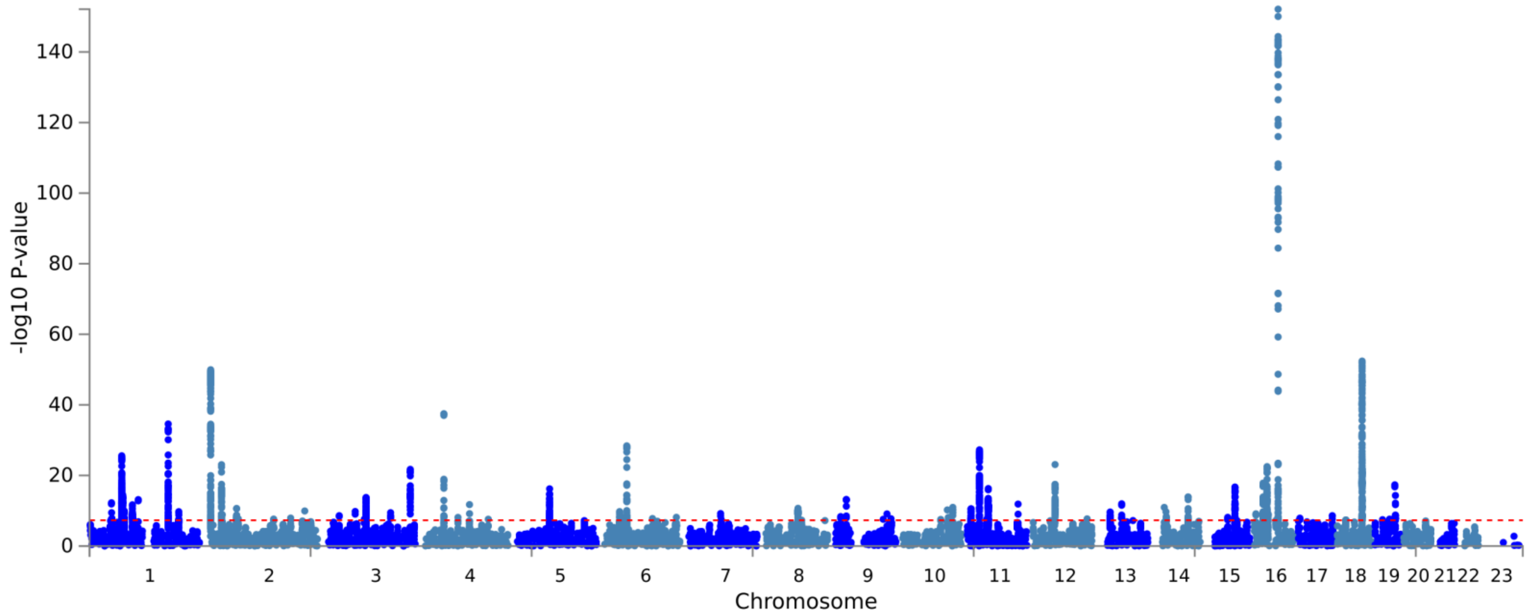
- SNPs that are within protein-coding gene regions
    - Default gene annotation window = 0Kb (would miss intergenic regulatory regions)
    - Options available in FUMA = 0, 5, 10, 15, 20Kb

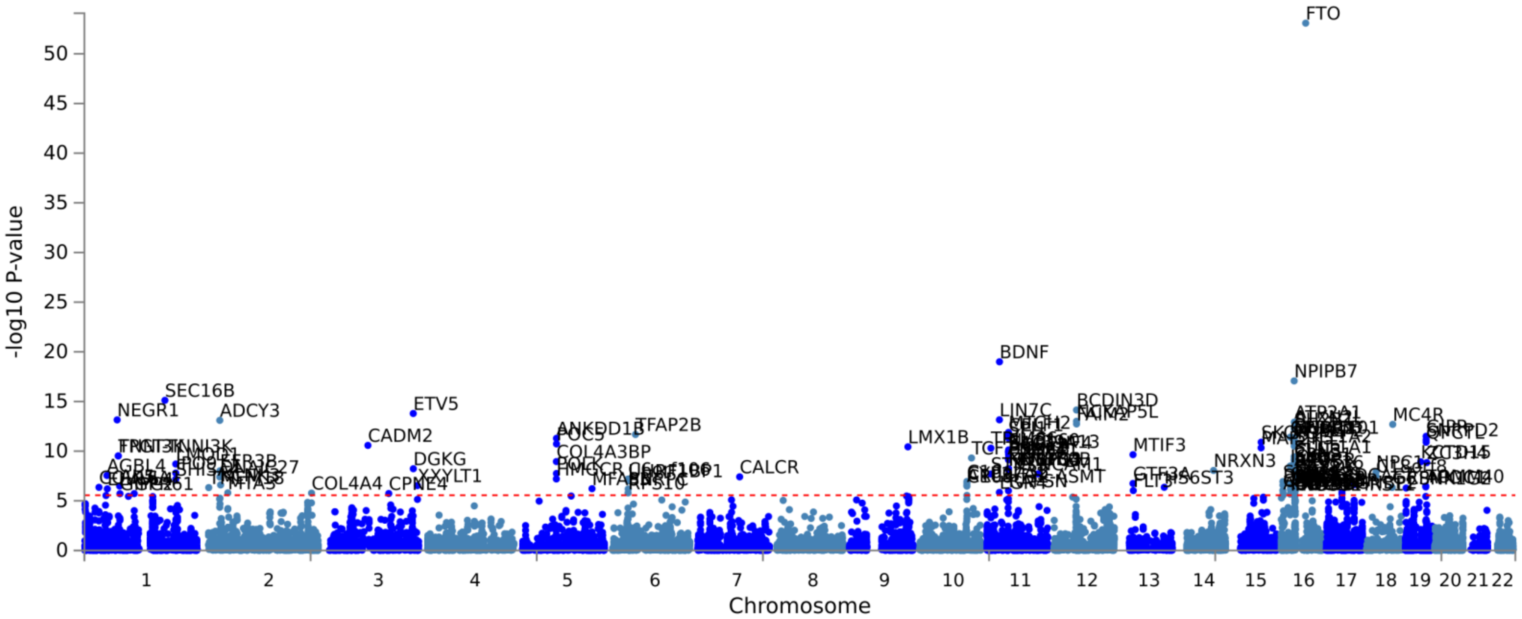
Step 2: Calculating gene p-value

- Multiple linear principal components regression model
- For each gene:
    - Project SNP matrix for the gene onto its principal components (uses 1000G phase 3 as reference data), removes redundant information and accounts for SNP-SNP LD
    - Uses PCs as predictors of phenotype in a linear regression model

$$Y = \alpha_{0g} + \overset{\text{Matrix of PCs}}{\overset{\downarrow}{X_g^*}} \underset{\text{genetic effect}}{\underset{\uparrow}{\alpha_g}} + \varepsilon_g$$

SNP-based vs MAGMA gene-based association for BMI





Gene-set analysis

Gene set - any group of genes that share a particular property e.g. sample pathway, same protein family etc

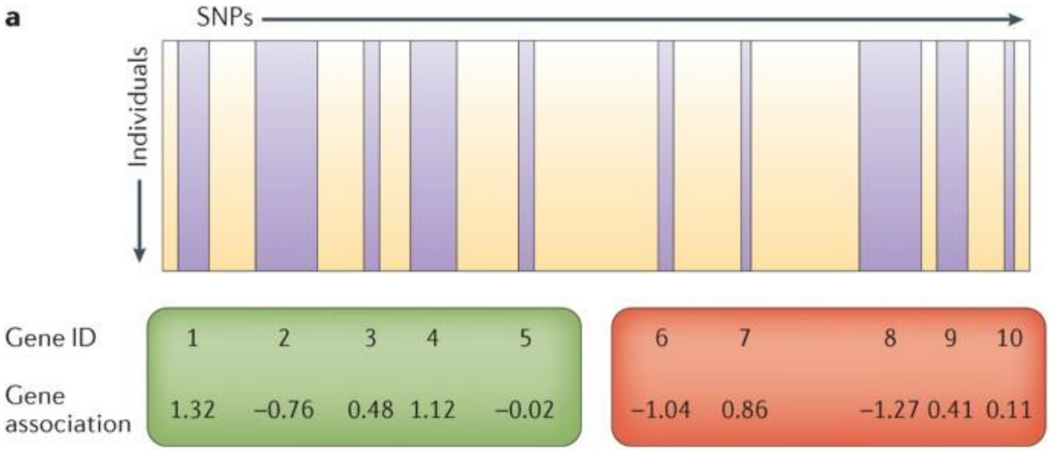
Gene set analysis - determine whether that property of the gene set has a role in the phenotype of interest.

1. Self-contained analysis:

- null hypothesis: none of the genes in the gene set are associated with phenotype.
- tests if genes in a gene-set are jointly associated with the phenotype of interest
- Only considers genes in the gene set

2. Competitive analysis:

- tests if genes in a gene-set more strongly associated with the phenotype than other genes
- Considers all genes in the data
- joint association of genes in the gene set is greater than the association of genes not in the gene set



Leeuw et al Nat Rev Gene 2016

Nature Reviews | Genetics

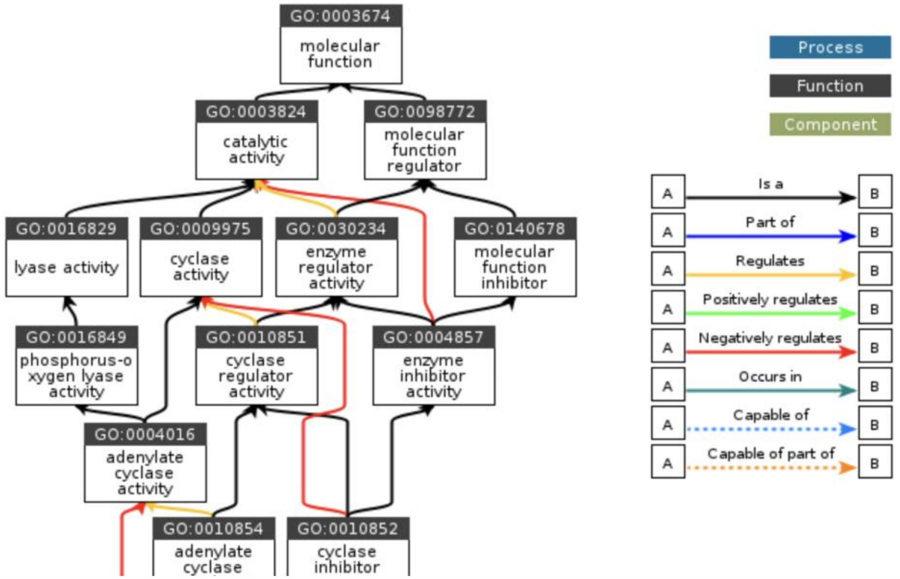
MAGMA gene-set analysis

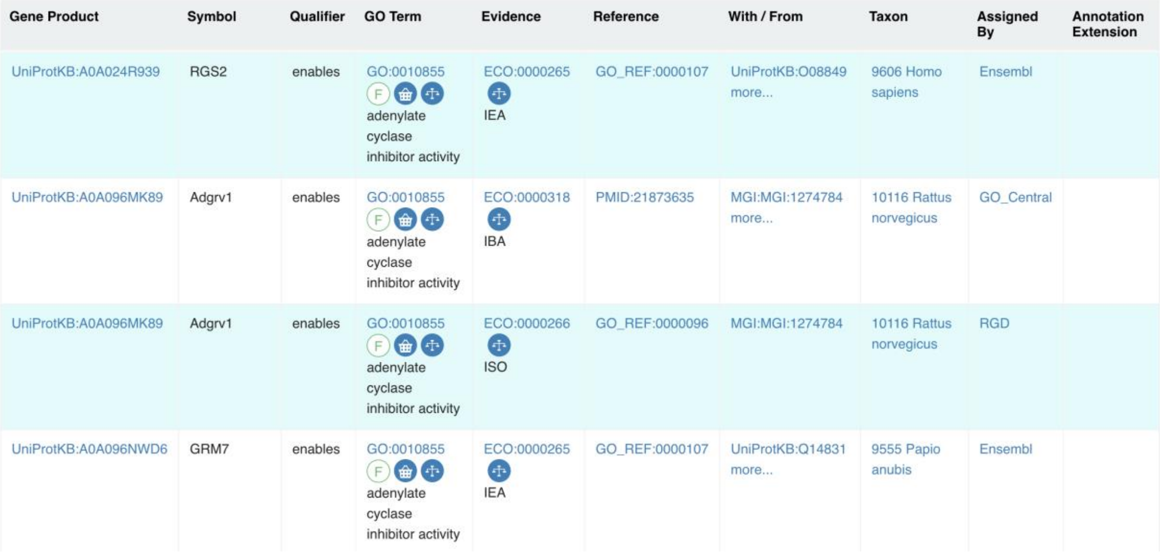
Competitive gene set analysis for 4728 curated gene sets (including canonical pathways) and 6166 GO terms

The Molecular Signatures Database (MSigDB) is a resource of annotated gene sets
https://www.gsea-msigdb.org/gsea/msigdb

- Online pathway databases: KEGG, Biocarta, Reactome, WikiPathwyas
- Biomedical literature
- Contributed by individual domain experts

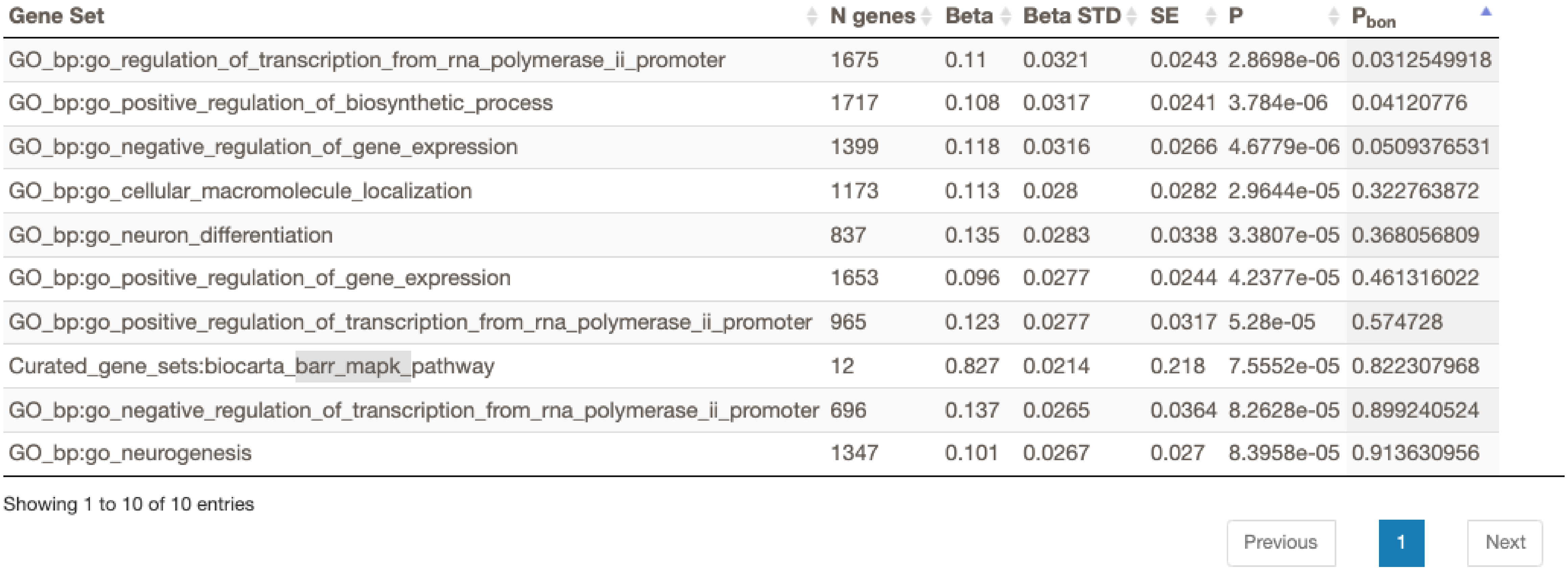
Gene Ontology - source of information on the functions of genes





Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon	Assigned By	Annotation Extension
UniProtKB:A0A024R939	RGS2	enables	GO:0010855 F adenylate cyclase inhibitor activity	ECO:0000265 IEA	GO_REF:0000107	UniProtKB:O08849 more...	9606 Homo sapiens	Ensembl	
UniProtKB:A0A096MK89	Adgrv1	enables	GO:0010855 F adenylate cyclase inhibitor activity	ECO:0000318 IBA	PMID:21873635	MGI:MGI:1274784 more...	10116 Rattus norvegicus	GO_Central	
UniProtKB:A0A096MK89	Adgrv1	enables	GO:0010855 F adenylate cyclase inhibitor activity	ECO:0000266 ISO	GO_REF:0000096	MGI:MGI:1274784	10116 Rattus norvegicus	RGD	
UniProtKB:A0A096NWD6	GRM7	enables	GO:0010855 F adenylate cyclase inhibitor activity	ECO:0000265 IEA	GO_REF:0000107	UniProtKB:Q14831 more...	9555 Papio anubis	Ensembl	

MAGMA gene-set analysis



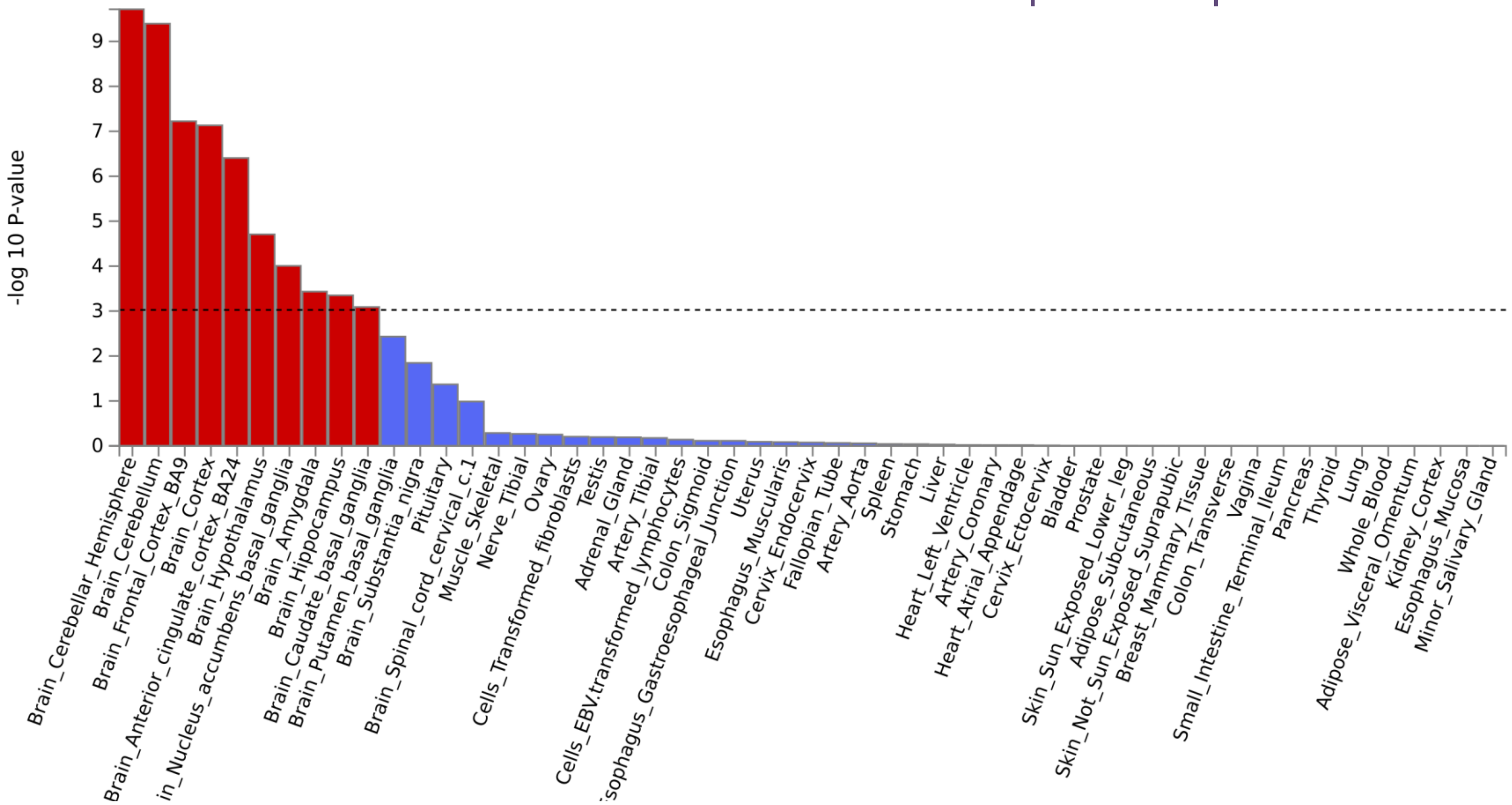
Gene Set	N genes	Beta	Beta STD	SE	P	P_{bon}
GO_bp:go_regulation_of_transcription_from_rna_polymerase_ii_promoter	1675	0.11	0.0321	0.0243	2.8698e-06	0.0312549918
GO_bp:go_positive_regulation_of_biosynthetic_process	1717	0.108	0.0317	0.0241	3.784e-06	0.04120776
GO_bp:go_negative_regulation_of_gene_expression	1399	0.118	0.0316	0.0266	4.6779e-06	0.0509376531
GO_bp:go_cellular_macromolecule_localization	1173	0.113	0.028	0.0282	2.9644e-05	0.322763872
GO_bp:go_neuron_differentiation	837	0.135	0.0283	0.0338	3.3807e-05	0.368056809
GO_bp:go_positive_regulation_of_gene_expression	1653	0.096	0.0277	0.0244	4.2377e-05	0.461316022
GO_bp:go_positive_regulation_of_transcription_from_rna_polymerase_ii_promoter	965	0.123	0.0277	0.0317	5.28e-05	0.574728
Curated_gene_sets:biocarta_barr_mapk_pathway	12	0.827	0.0214	0.218	7.5552e-05	0.822307968
GO_bp:go_negative_regulation_of_transcription_from_rna_polymerase_ii_promoter	696	0.137	0.0265	0.0364	8.2628e-05	0.899240524
GO_bp:go_neurogenesis	1347	0.101	0.0267	0.027	8.3958e-05	0.913630956

Showing 1 to 10 of 10 entries

Are genes in a gene-set more strongly associated with the phenotype of interest than other genes.

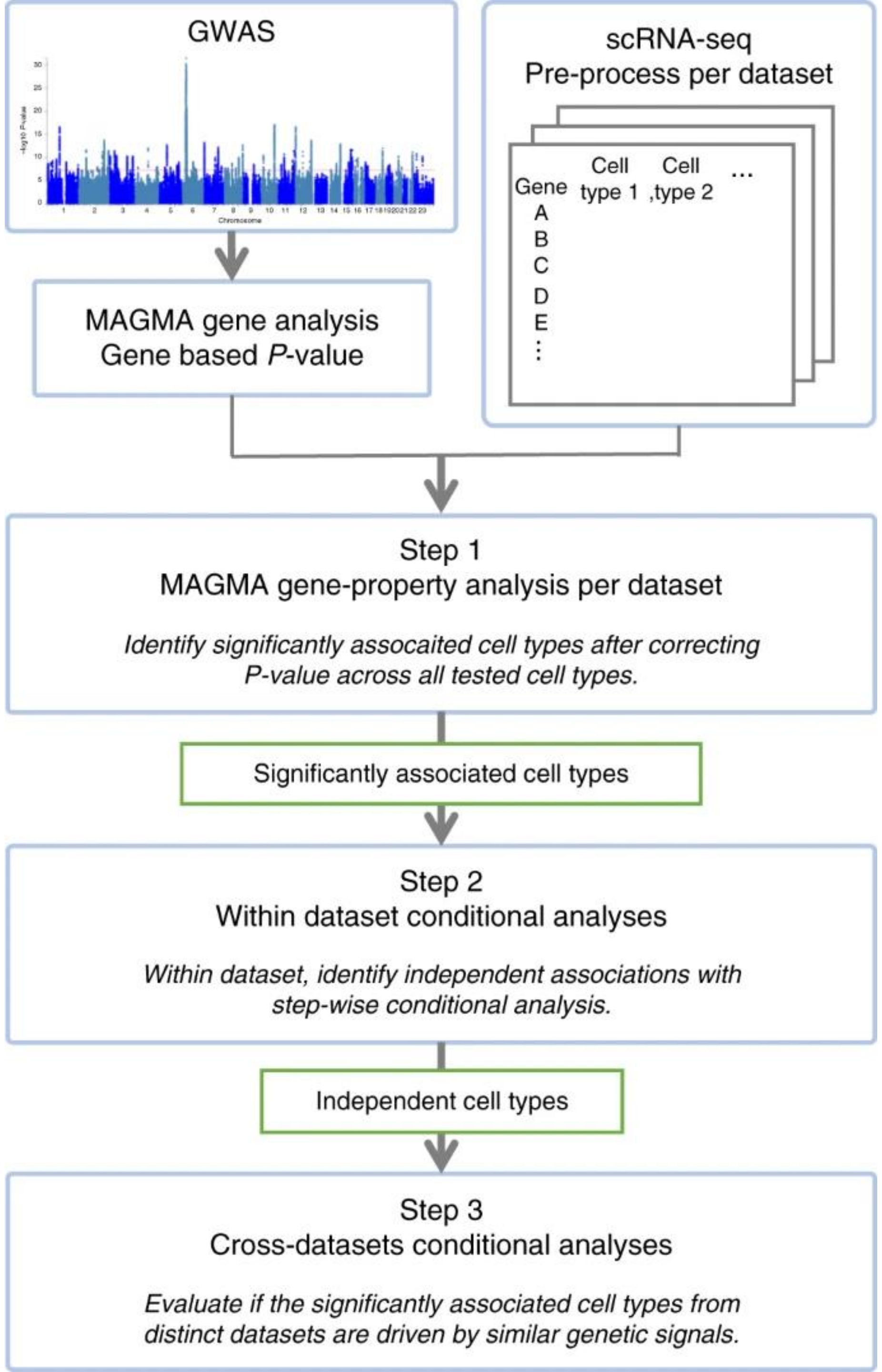
MAGMA tissue expression analysis

Do the genes most strongly associated with the phenotype have tissue-specific expression?



One-sided test if $\beta_{\varepsilon} > 0$

i.e. testing the positive relationship between tissue specificity and genetic association of genes.



GWAS

scRNA-seq
Pre-process per dataset

Gene	Cell type 1	Cell ,type 2	...
A			
B			
C			
D			
E			
⋮			

MAGMA gene analysis
Gene based *P*-value

Step 1
MAGMA gene-property analysis per dataset

Identify significantly assocaited cell types after correcting P-value across all tested cell types.

Significantly associated cell types

Step 2
Within dataset conditional analyses

Within dataset, identify independent associations with step-wise conditional analysis.

Independent cell types

Step 3
Cross-datasets conditional analyses

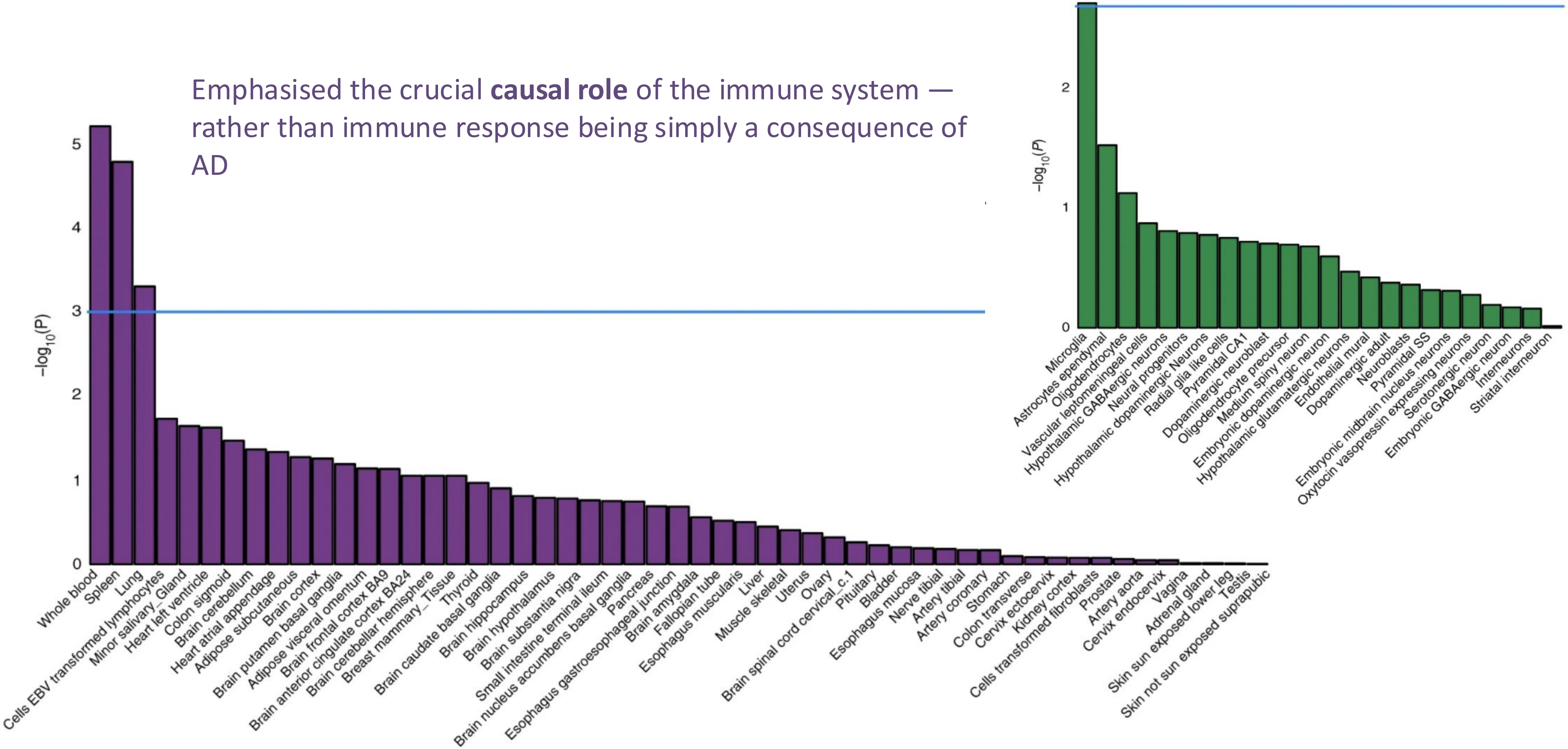
Evaluate if the significantly associated cell types from distinct datasets are driven by similar genetic signals.

FUMA Output

- List of prioritized variants and genes relevant to the trait/disease
- Biological pathways or functions relevant to the trait/disease
- Tissues/cell types relevant for trait/disease

Example: Role of microglia in Alzheimer’s disease

Emphasised the crucial **causal role** of the immune system — rather than immune response being simply a consequence of AD


$-\log_{10}(P)$
Whole blood
Spleen
Lung
Cells EBV transformed lymphocytes
Minor salivary_Gland
Heart left ventricle
Colon sigmoid
Brain cerebellum
Heart atrial appendage
Adipose subcutaneous
Brain cortex
Brain putamen basal ganglia
Adipose visceral omentum
Brain frontal cortex BA9
Brain anterior cingulate cortex BA24
Brain cerebellar hemisphere
Breast mammary_Tissue
Thyroid
Brain caudate basal ganglia
Brain hippocampus
Brain hypothalamus
Brain substantia nigra
Small intestine terminal ileum
Brain nucleus accumbens basal ganglia
Pancreas
Esophagus gastroesophageal junction
Brain amygdala
Fallopian tube
Esophagus muscularis
Liver
Muscle skeletal
Uterus
Ovary
Brain spinal cord cervical_c.1
Pituitary
Bladder
Esophagus mucosa
Nerve tibial
Artery tibial
Artery coronary
Stomach
Colon transverse
Cervix ectocervix
Kidney cortex
Cells transformed fibroblasts
Prostate
Artery aorta
Cervix endocervix
Vagina
Adrenal gland
Skin sun exposed lower leg
Testis
Skin not sun exposed suprapubic
$-\log_{10}(P)$
Microglia
Astrocytes ependymal
Oligodendrocytes
Vascular leptomeningeal cells
Hypothalamic GABAergic neurons
Neural progenitors
Hypothalamic dopaminergic Neurons
Radial glia like cells
Pyramidal CA1
Dopaminergic neuroblast
Oligodendrocyte precursor
Medium spiny neuron
Embryonic dopaminergic neuron
Hypothalamic glutamatergic neurons
Endothelial mural
Dopaminergic adult
Neuroblasts
Pyramidal SS
Embryonic midbrain nucleus neurons
Oxytocin vasopressin expressing neurons
Serotonergic neuron
Embryonic GABAergic neuron
Interneurons
Striatal interneuron