The University
Of Queensland
AUSTRALIA

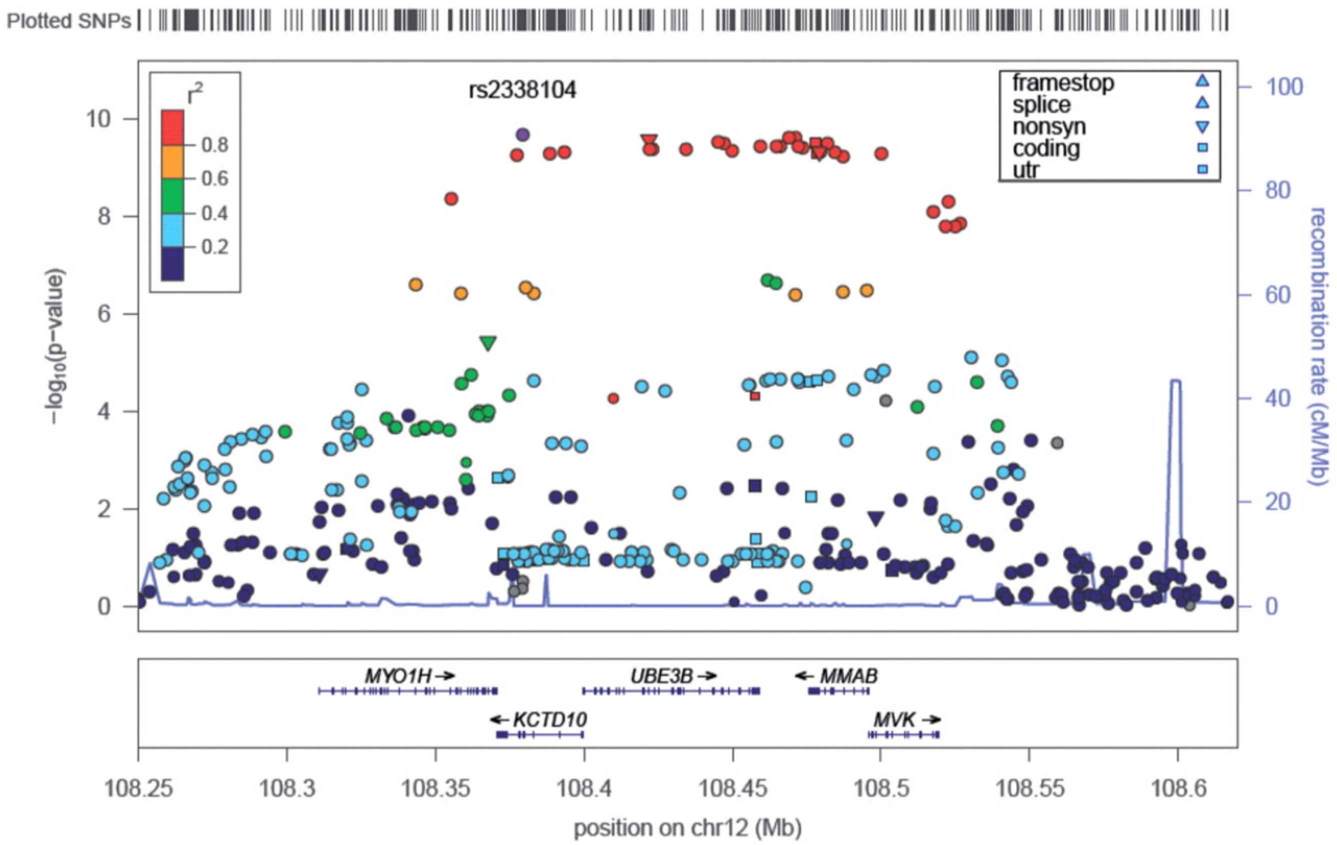
Functional annotation of GWAS summary data using FUMA

Biological interpretation of GWAS loci

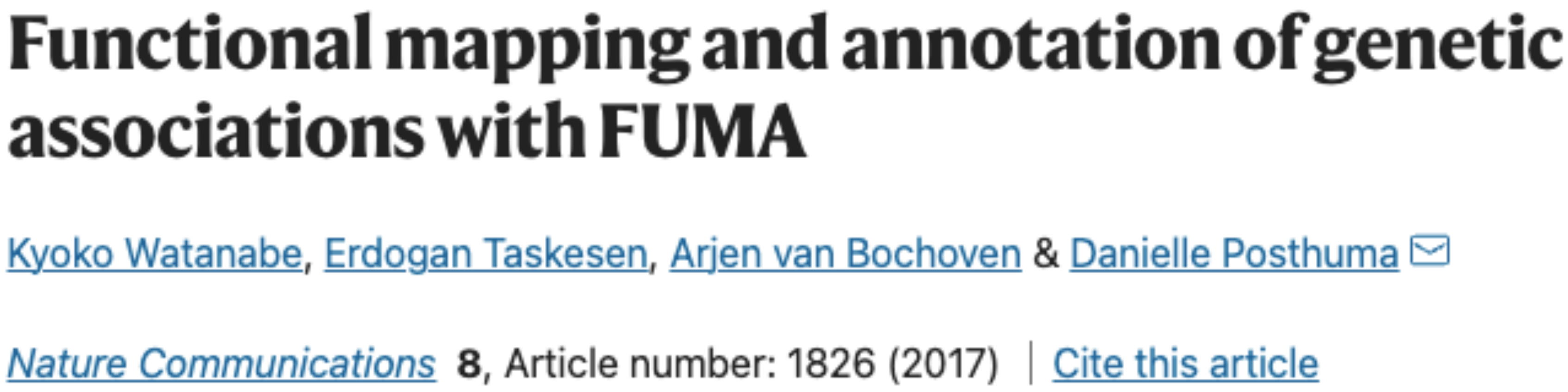
- GWAS hits span a genomic region characterized by multiple correlated SNPs
  - May cover multiple closely located genes
  - May be in intergenic or non-coding regions

- Identifying causal variant and gene often not possible based on p-values alone

- Requires integration of LD information and functional consequences



HDL cholesterol-associated region near the MMAB gene (Kathiresan et al Nature Genetics 2009)



Functional mapping and annotation of genetic associations with FUMA

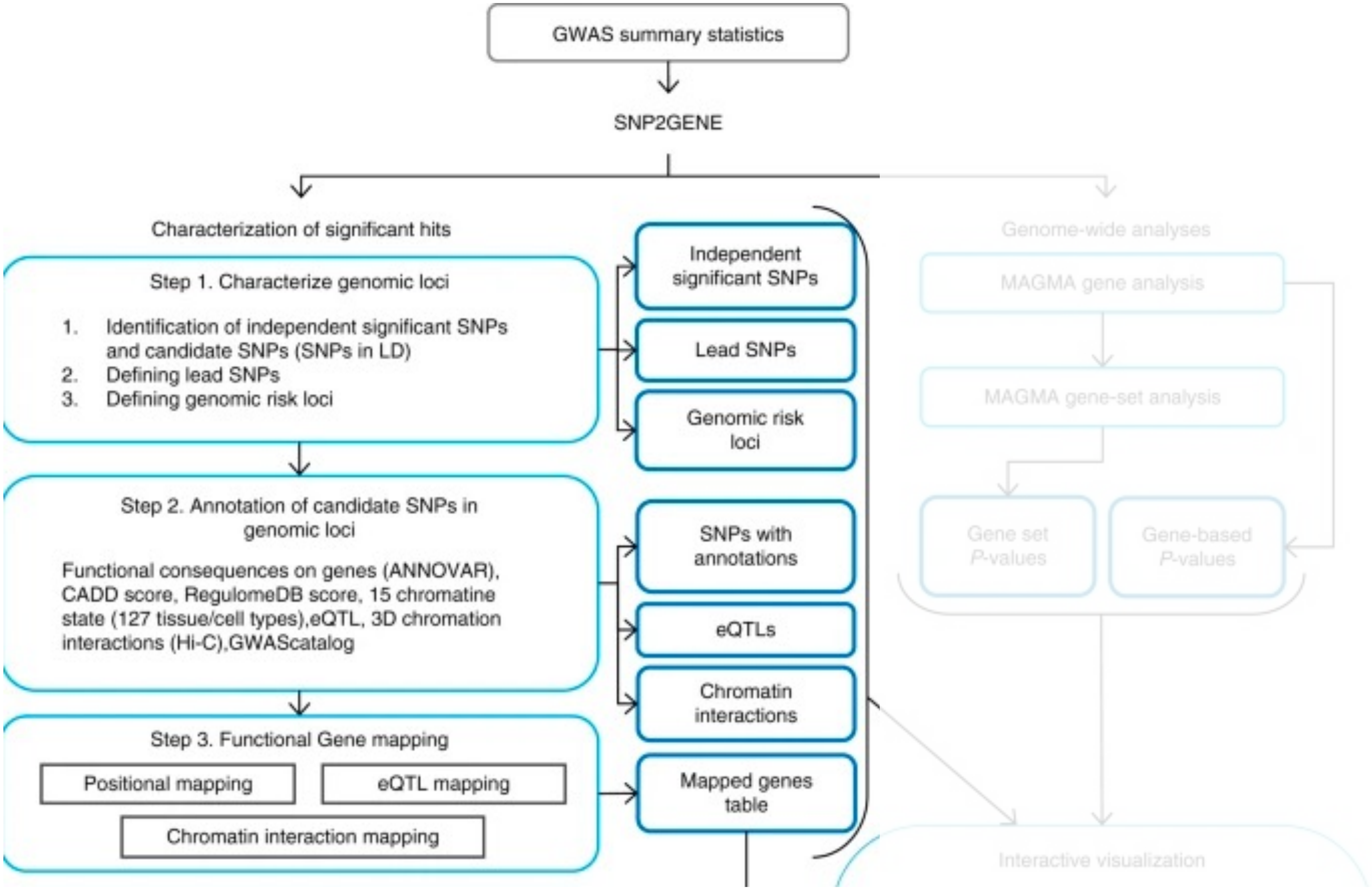
Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven & Danielle Posthuma

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

Incorporates 18 biological repositories and tools to process GWAS summary data:

- SNP2GENE: mapping of SNPs to genes based on positional, eQTL and chromatin interaction

- GENE2FUNC: biological mechanisms of prioritized genes



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

Independent and candidate SNPs

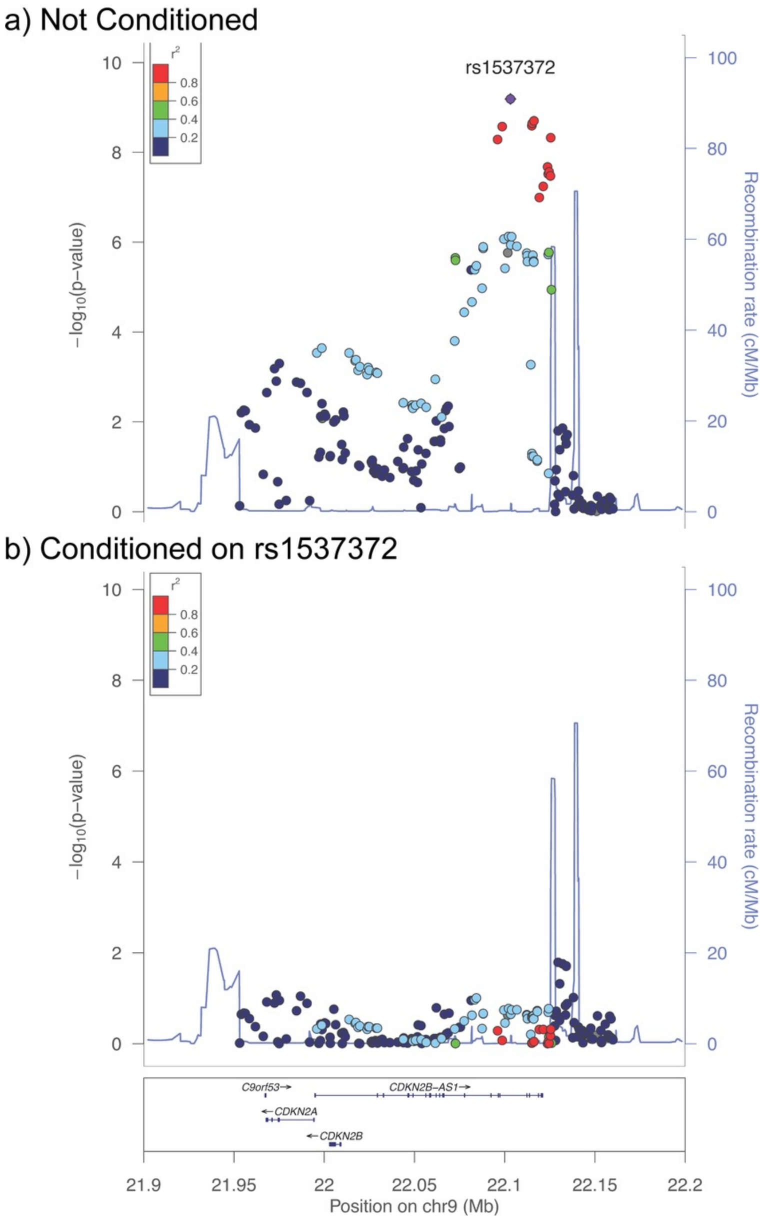
1. Independent significant SNPs

- LD Clumping
- SNPs with *P*-value < 5e-8 and independent from each other at $r^2 < 0.6$

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 LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs.

4. Functional consequence of candidate SNPs on genes using ANNOVAR



a) Not Conditioned

b) Conditioned on rs1537372

Integration of Functional Resources

Combined Annotation Dependent Depletion (CADD)

Encyclopedia of DNA Elements (ENCODE)

Roadmap Epigenomics Project

Chromatin interaction information

The Genotype-Tissue Expression (GTEx) and other eQTL data

CADD

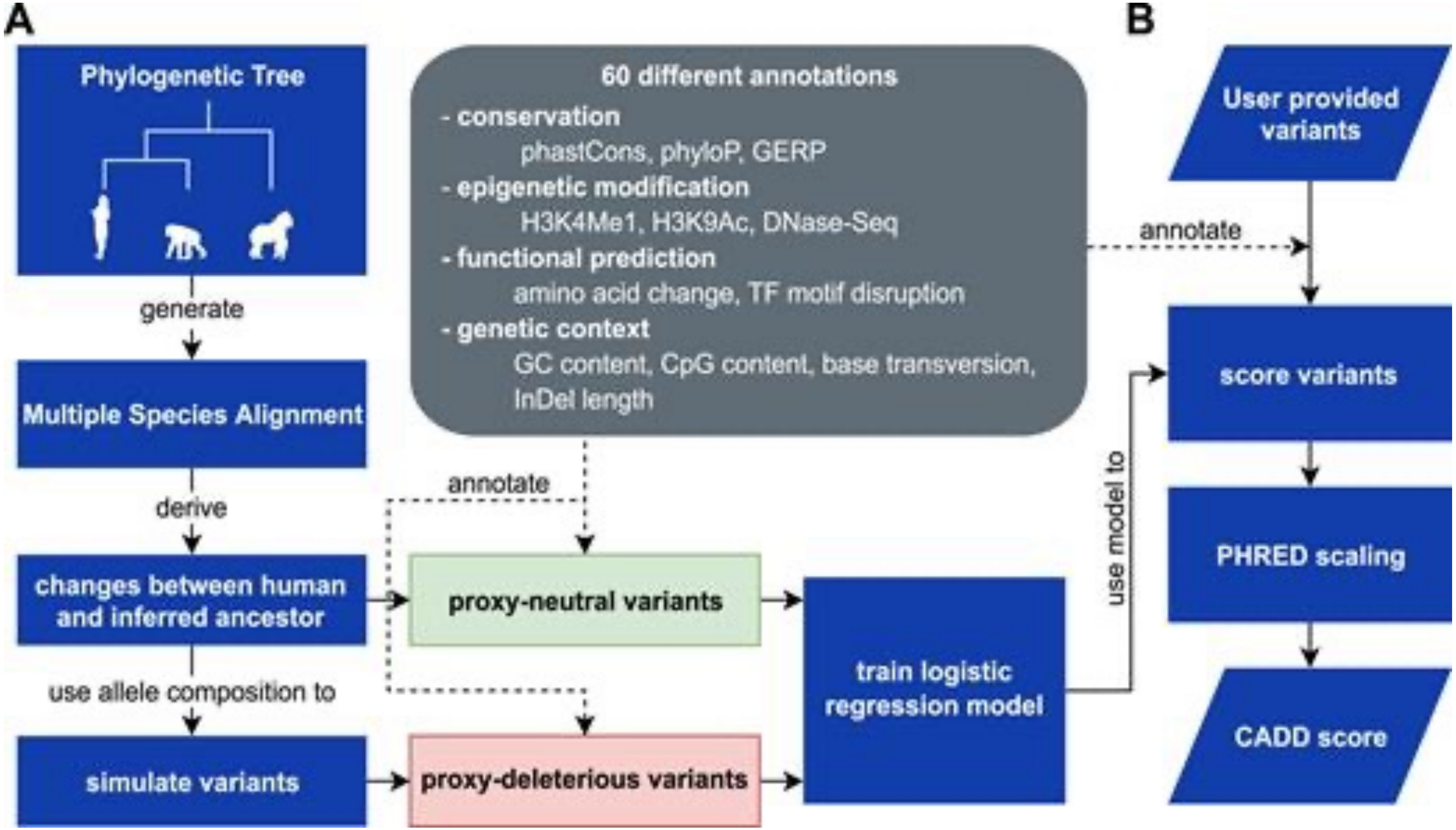
- CADD score - a measure of variant deleteriousness (reduce organismal fitness) based on predictive genomics features (Kircher et al Nature Genetics 2014).

- **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 that best separates these two sets of variants (assumption: deleterious variants will depleted in observed variants compared to simulated 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

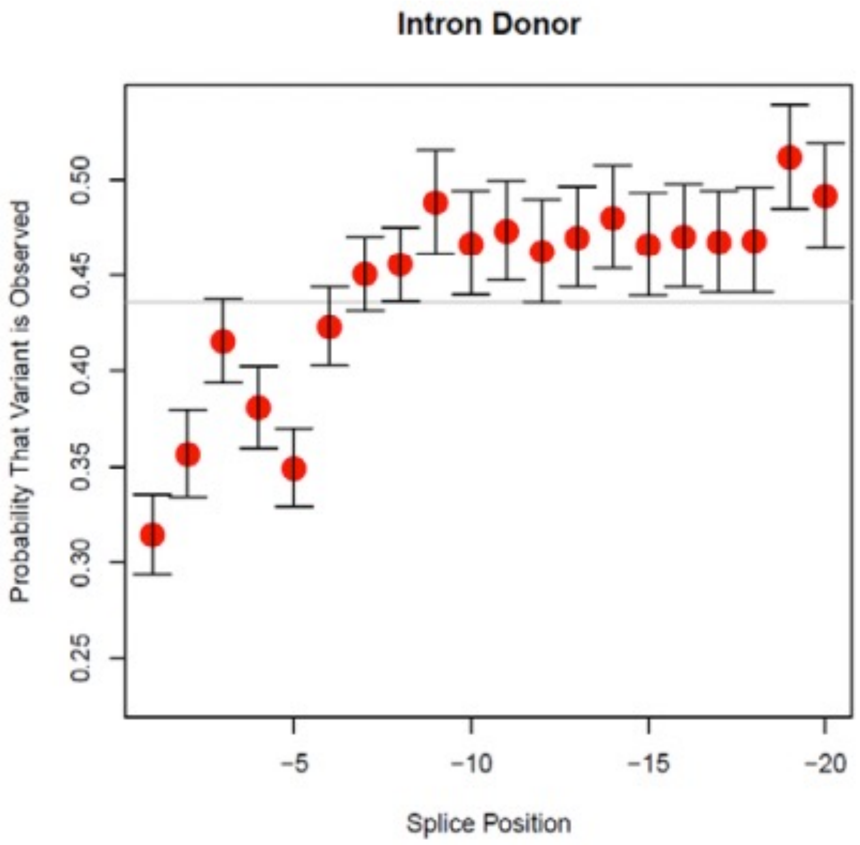
Raw c-score: the probability of a variant coming from the simulated vs observed set based on it’s annotation profile

Higher the value more likely to be deleterious.

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



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

Process:

- Formaldehyde to covalently link DNA regions that are in close spatial proximity in nucleus
- DNA is cleaved by restriction digestion and DNA ends are filled in with biotinylated nucleotides.
- DNA is then ligated to form hybrid DNA molecules, each corresponding to an interaction event of a pair of loci.
- Fragmented and sequenced, then mapped to genome

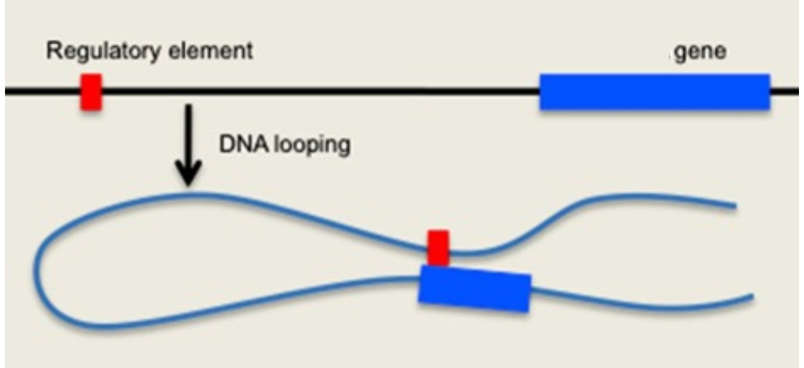
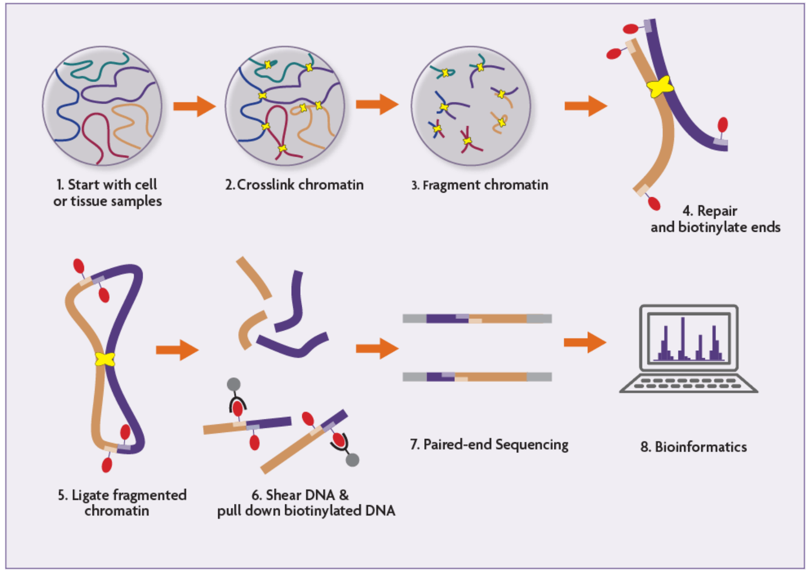


Figure DOI: 10.3389/fnmol.2013.00032

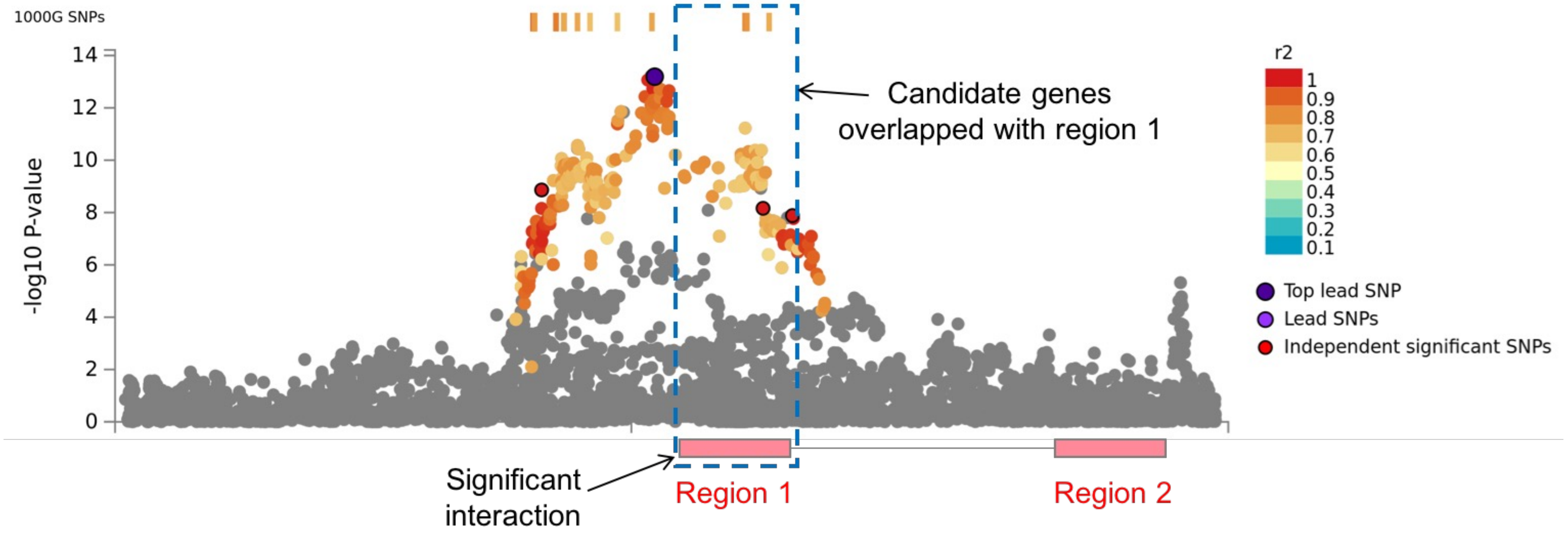

1. Start with cell or tissue samples
2. Crosslink chromatin
3. Fragment chromatin
4. Repair and biotinylate ends
5. Ligate fragmented chromatin
6. Shear DNA & pull down biotinylated DNA
7. Paired-end Sequencing
8. Bioinformatics

Source: https://www.activemotif.com.cn/

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



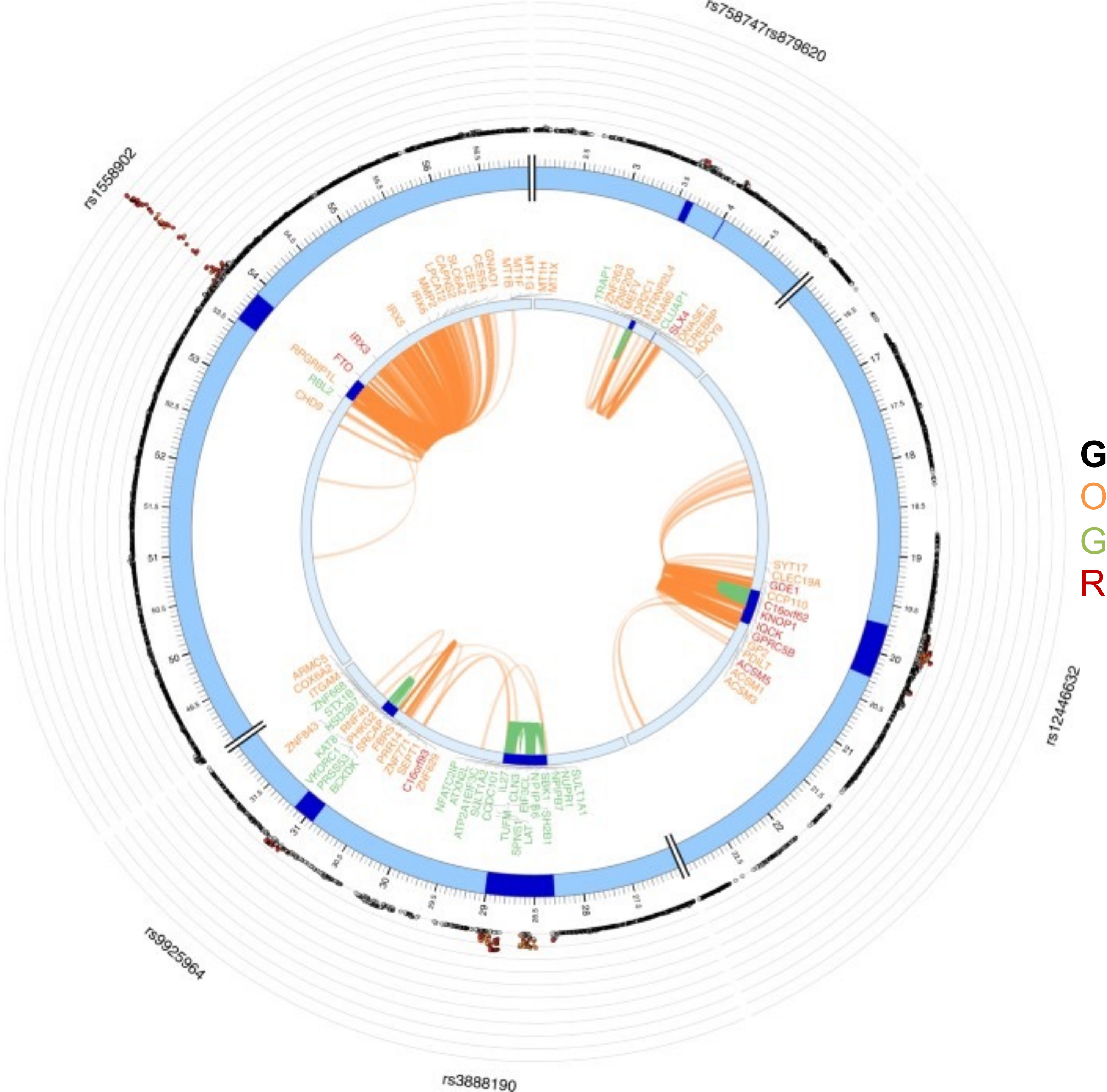
Candidate genes overlapped with region 1

Significant interaction

Region 1

Region 2

Chromatin interactions and eQTLS of a BMI risk locus on chr16



Genes

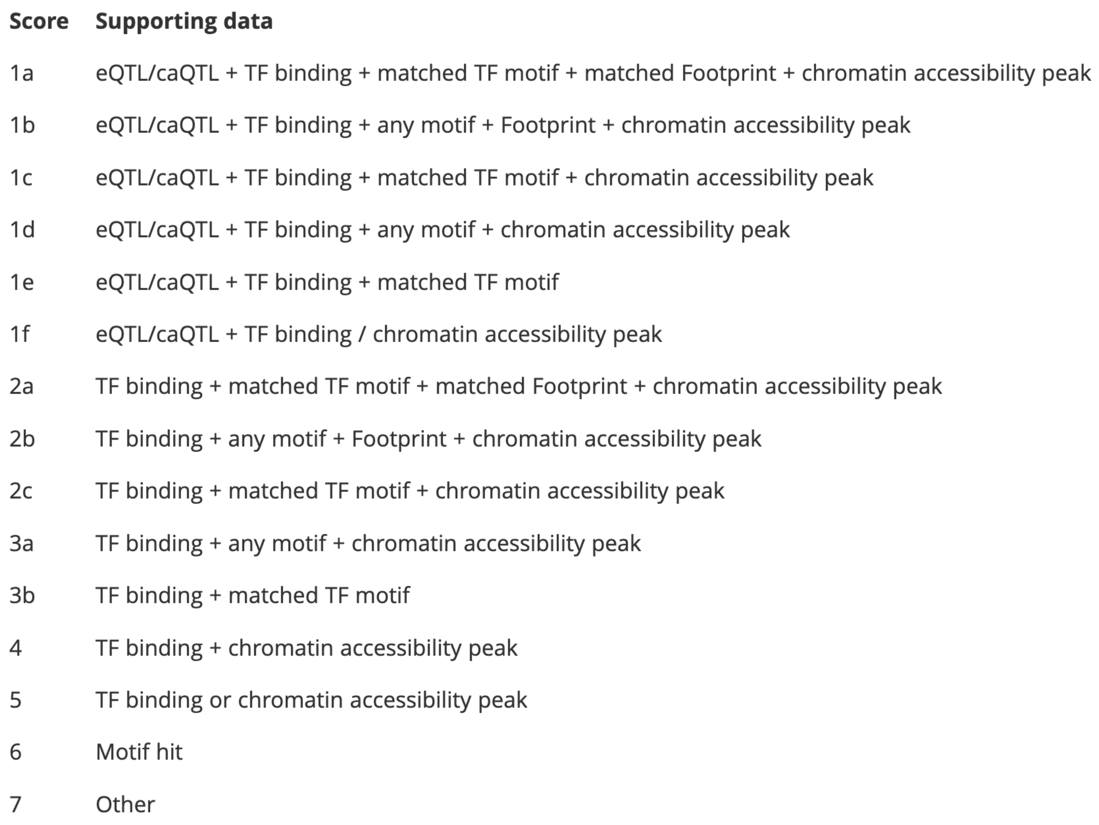
Orange: mapped by eQTL data

Green: mapped by HiC data

Red: mapped by both

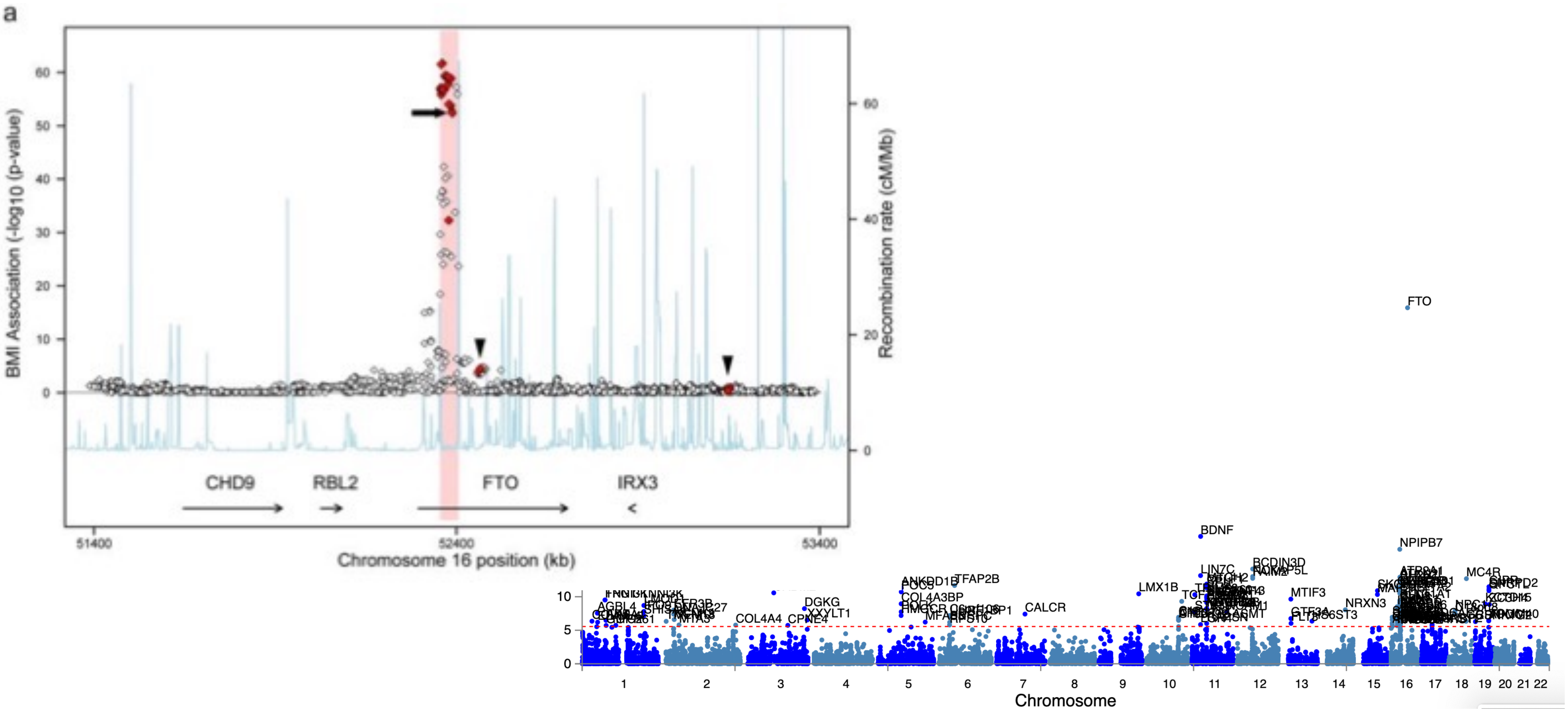
RegulomeDB

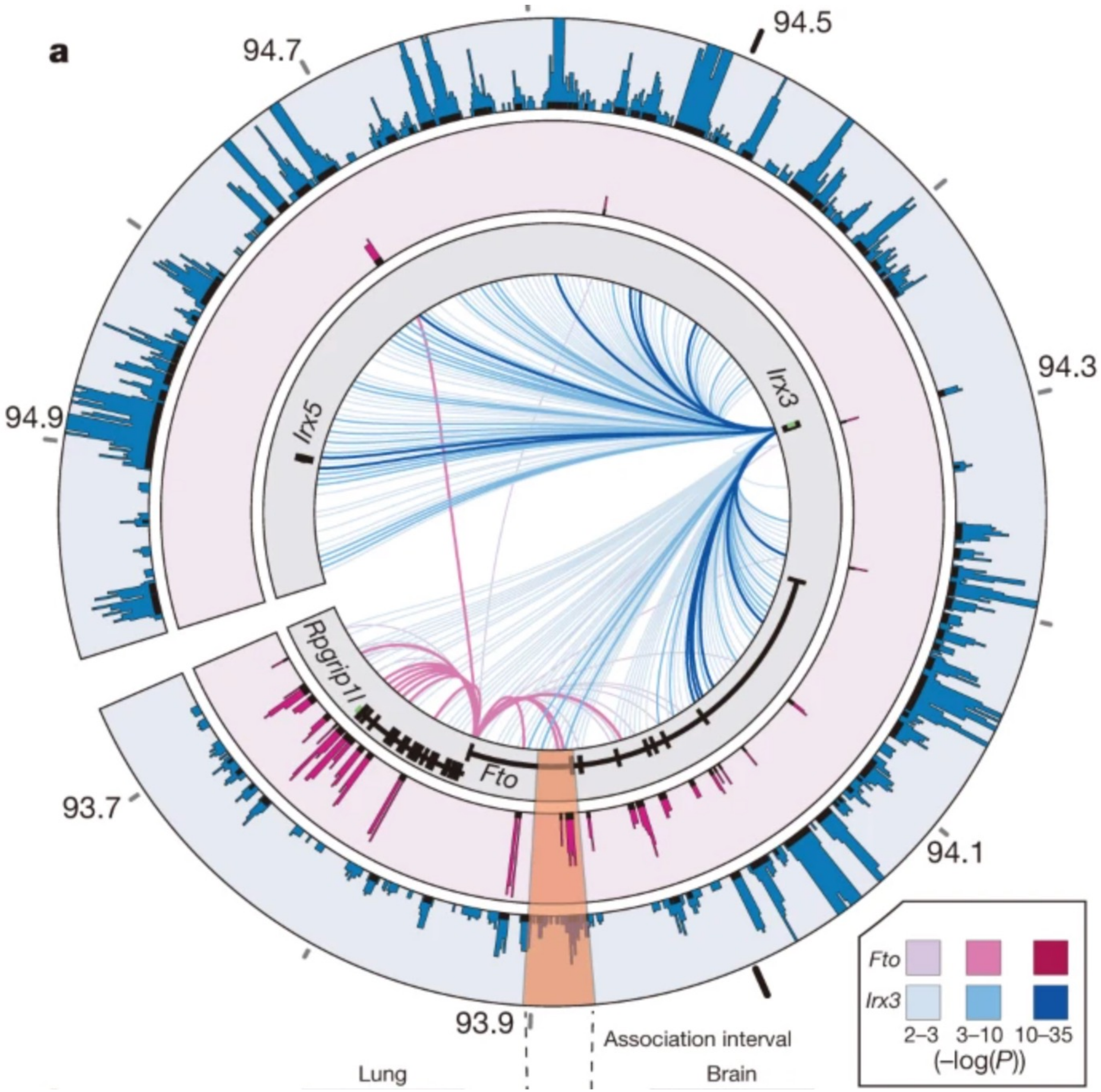
- Intersects candidate SNPs with known functionally-active regions identified from functional genomic assays e.g. TF ChiP-seq (TF-binding regions), DNAase-seq (open chromatin regions)
- Scores functional consequence of each SNP based on strength of evidence



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 example - *FTO*





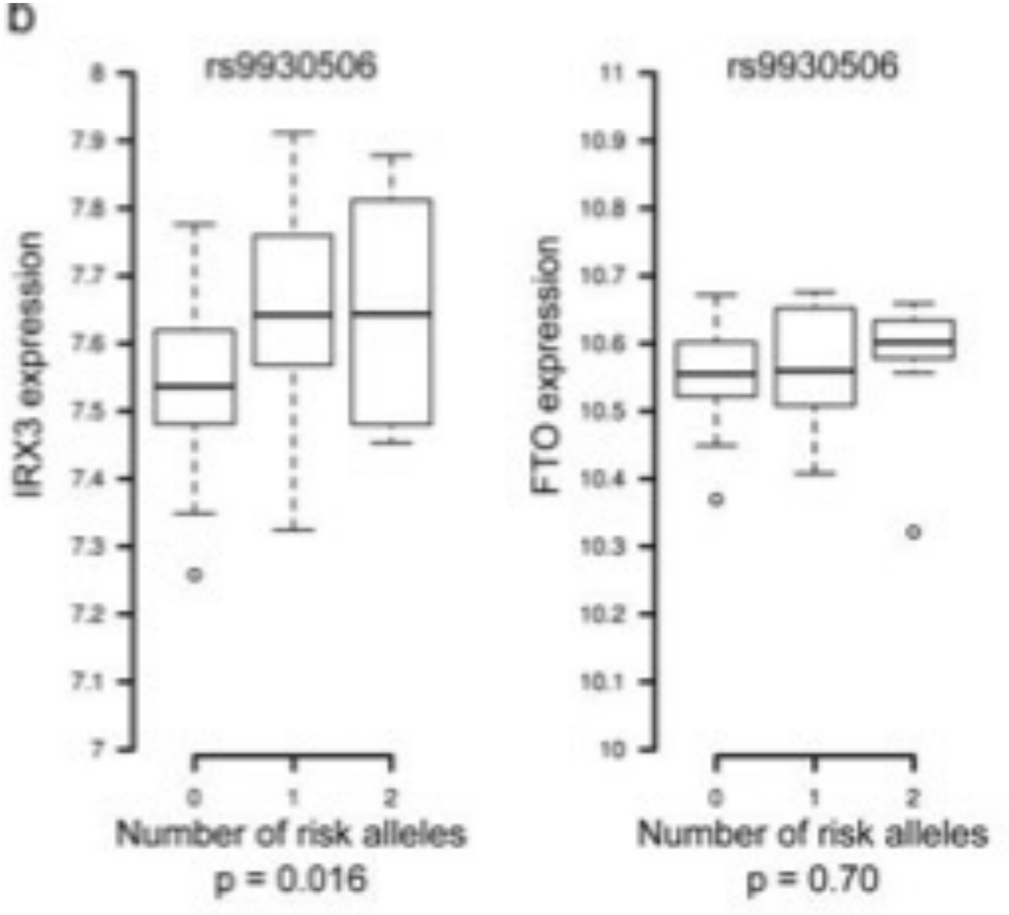
Smemo et al Nature 2014

The *Fto* promoter chiefly participates in genomic interactions proximal to the gene promoter.

Interaction between Fto promoter and GWAS region in mouse embryos but not in adult mouse brains

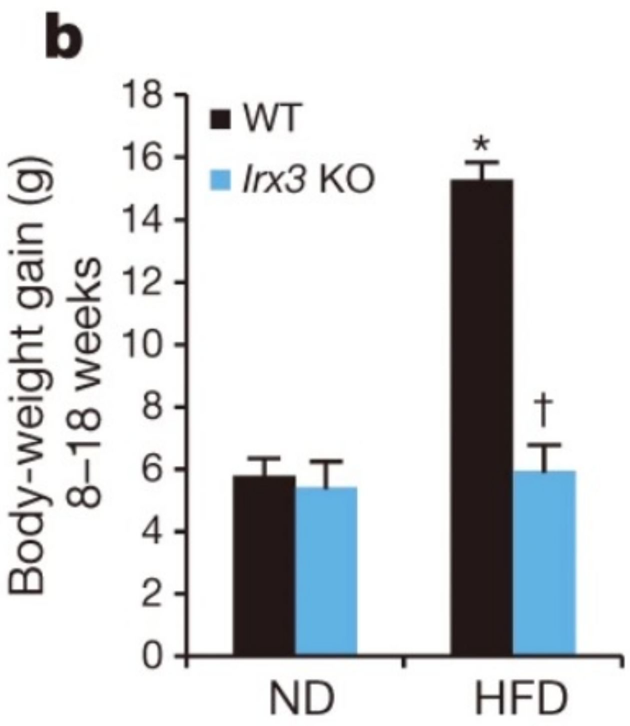
Promoter of *Irx3* participates in numerous long-range interactions, including with the GWAS region in both mouse embryo and adult mouse brain, as well as MCF-7 cells and zebrafish embryos

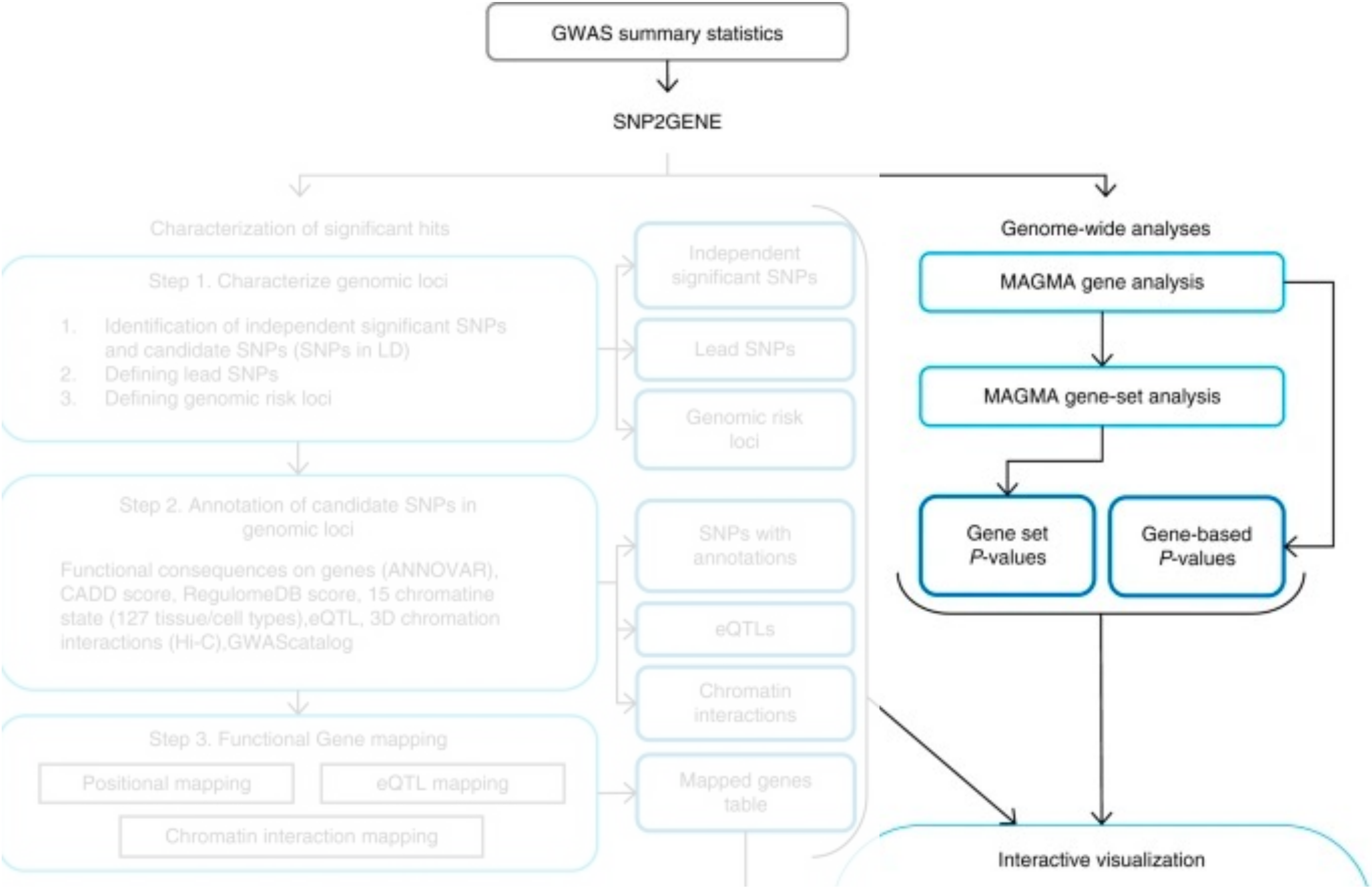
Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*



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 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

Gene-based test

- 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 χ2-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

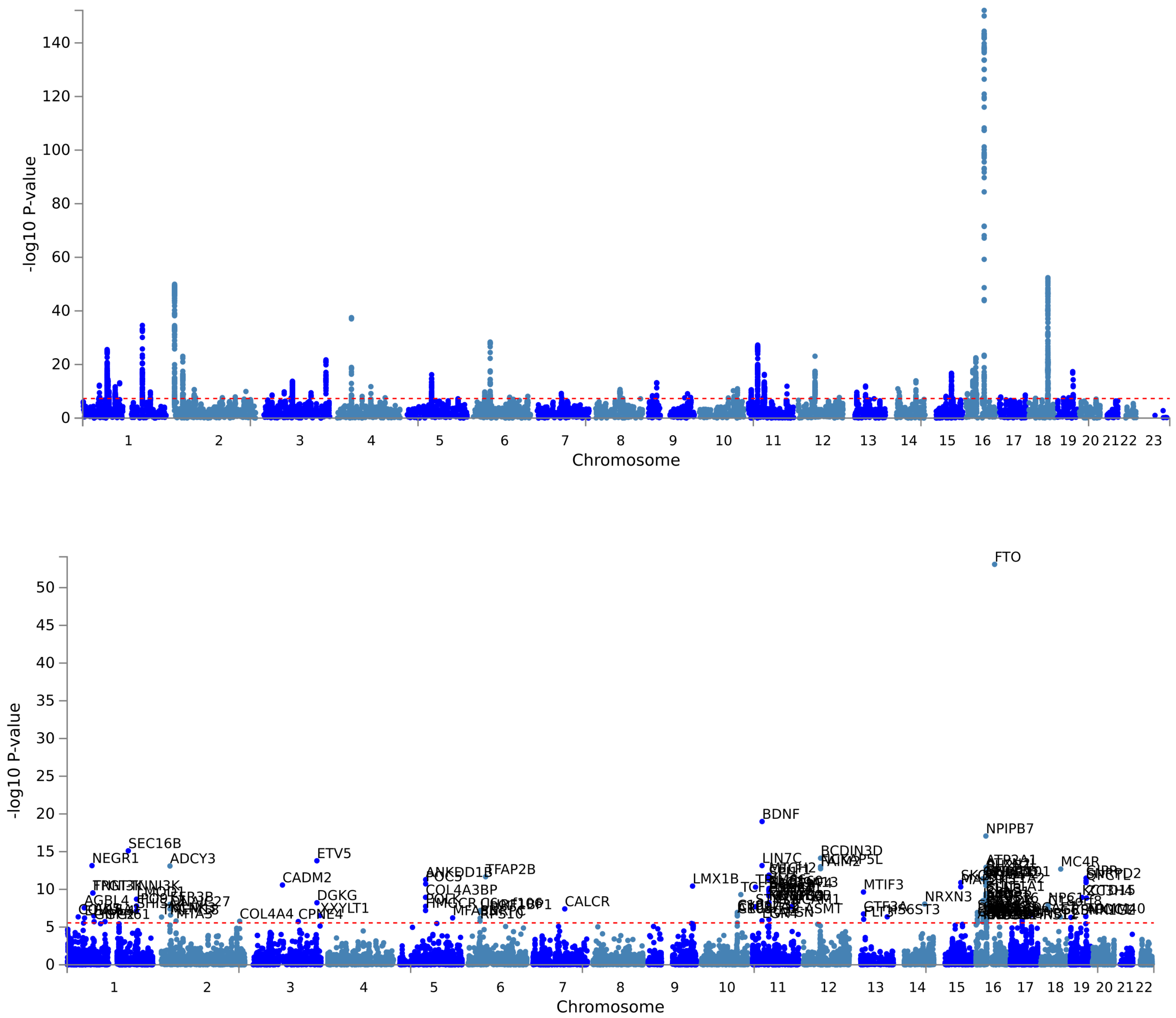
Step 1: Mapping SNPs to gene

- 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

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



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.

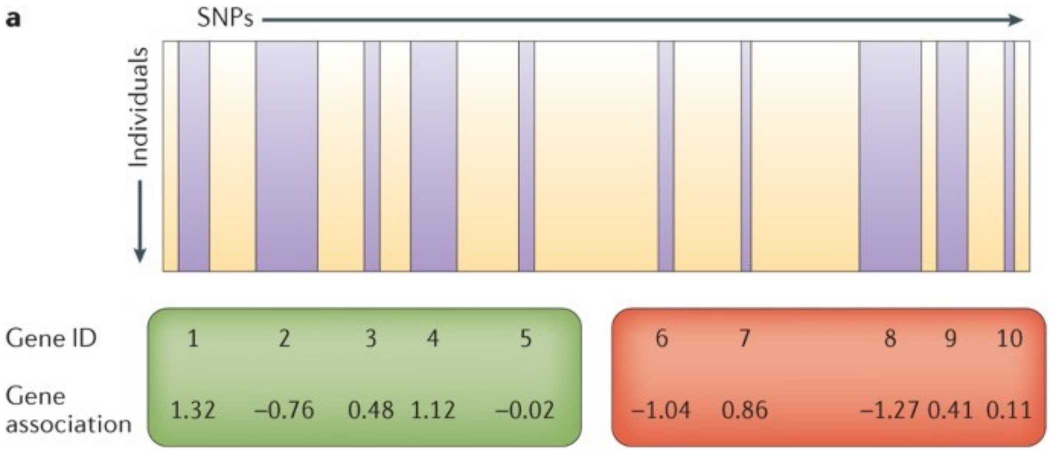
Two approaches

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 gene in gene 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

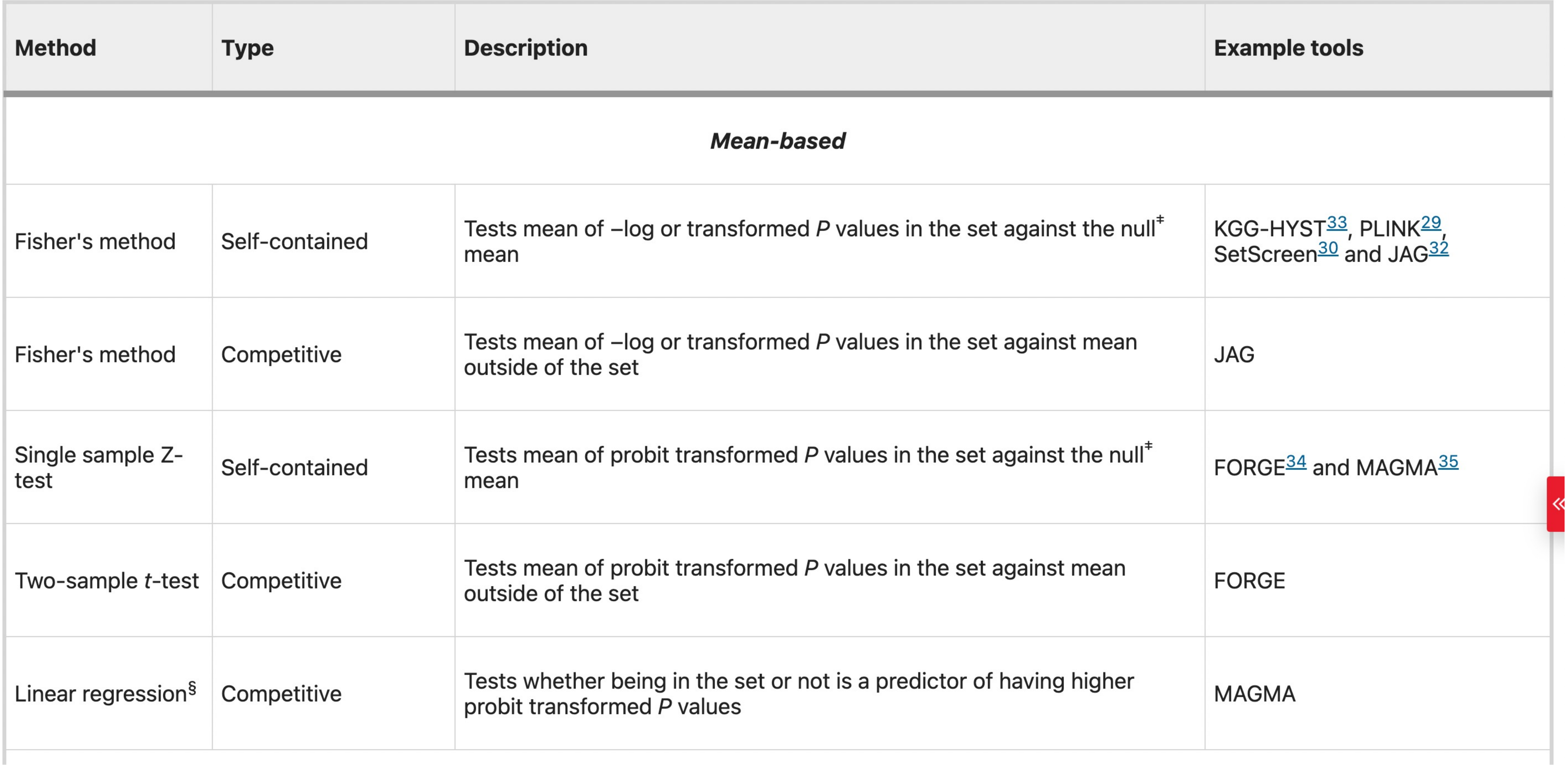
Gene-set analysis

Type of gene-based test-statistic for competitive analysis

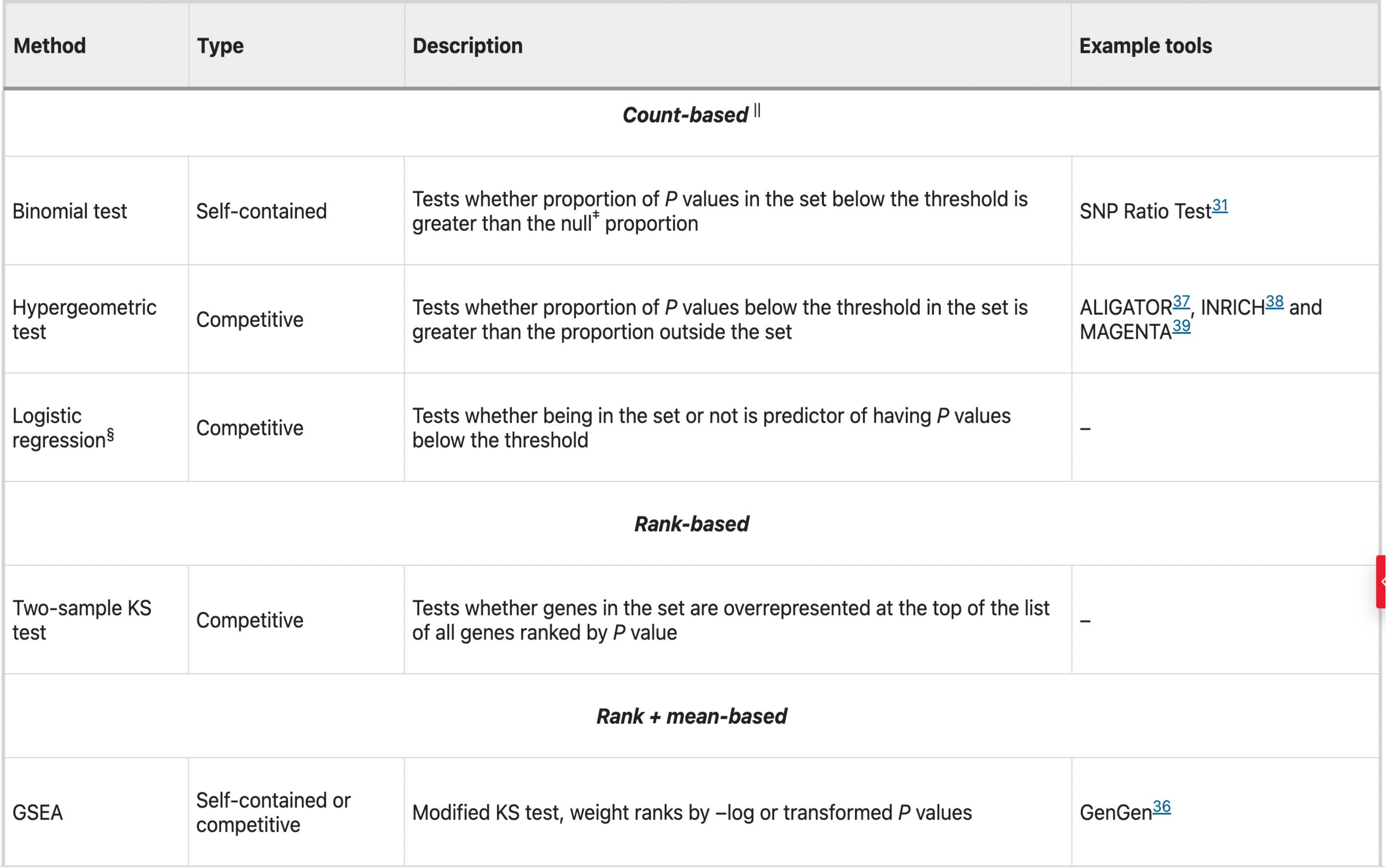
- **mean-based**, using the (possibly weighted) mean or sum of the gene-association scores;

- **count-based**, classifying genes as 'significant' or 'not significant' by applying a threshold to the gene-association scores and using the number of 'significant' genes in the gene set as a test statistic;

- **rank-based**, ranking the genes on their gene-association score and computing overrepresentation of the gene-set genes at the top of that ranking.

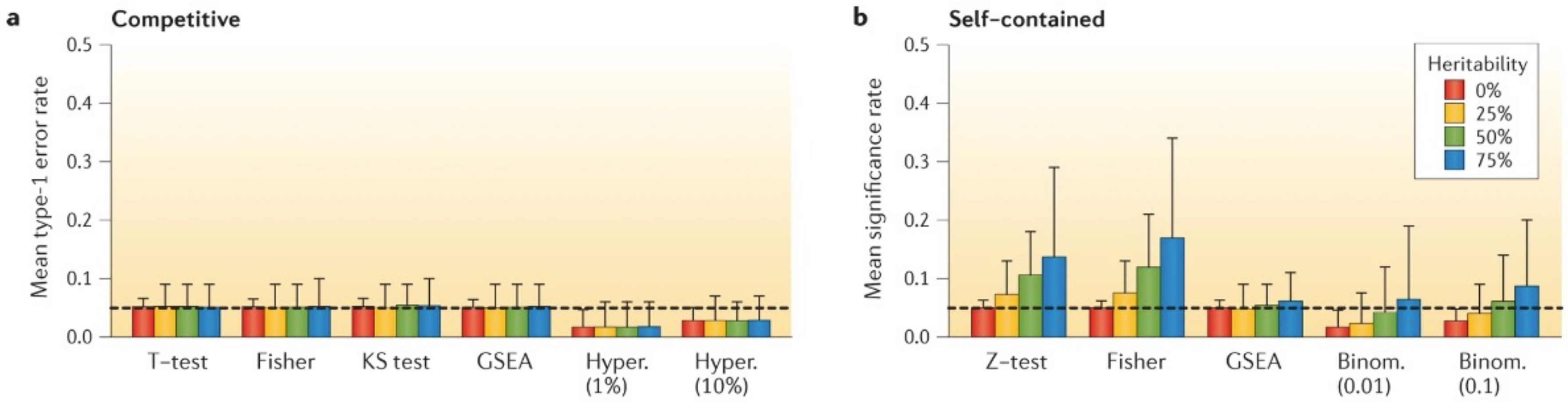


Method	Type	Description	Example tools
Mean-based			
Fisher's method	Self-contained	Tests mean of –log or transformed *P* values in the set against the null[‡] mean	KGG-HYST[33], PLINK[29], SetScreen[30] and JAG[32]
Fisher's method	Competitive	Tests mean of –log or transformed *P* values in the set against mean outside of the set	JAG
Single sample Z-test	Self-contained	Tests mean of probit transformed *P* values in the set against the null[‡] mean	FORGE[34] and MAGMA[35]
Two-sample *t*-test	Competitive	Tests mean of probit transformed *P* values in the set against mean outside of the set	FORGE
Linear regression[§]	Competitive	Tests whether being in the set or not is a predictor of having higher probit transformed *P* values	MAGMA



Method	Type	Description	Example tools
Count-based [‖]			
Binomial test	Self-contained	Tests whether proportion of *P* values in the set below the threshold is greater than the null[‡] proportion	SNP Ratio Test[31]
Hypergeometric test	Competitive	Tests whether proportion of *P* values below the threshold in the set is greater than the proportion outside the set	ALIGATOR[37], INRICH[38] and MAGENTA[39]
Logistic regression[§]	Competitive	Tests whether being in the set or not is predictor of having *P* values below the threshold	–
Rank-based			
Two-sample KS test	Competitive	Tests whether genes in the set are overrepresented at the top of the list of all genes ranked by *P* value	–
Rank + mean-based			
GSEA	Self-contained or competitive	Modified KS test, weight ranks by –log or transformed *P* values	GenGen[36]

From: The statistical properties of gene-set analysis

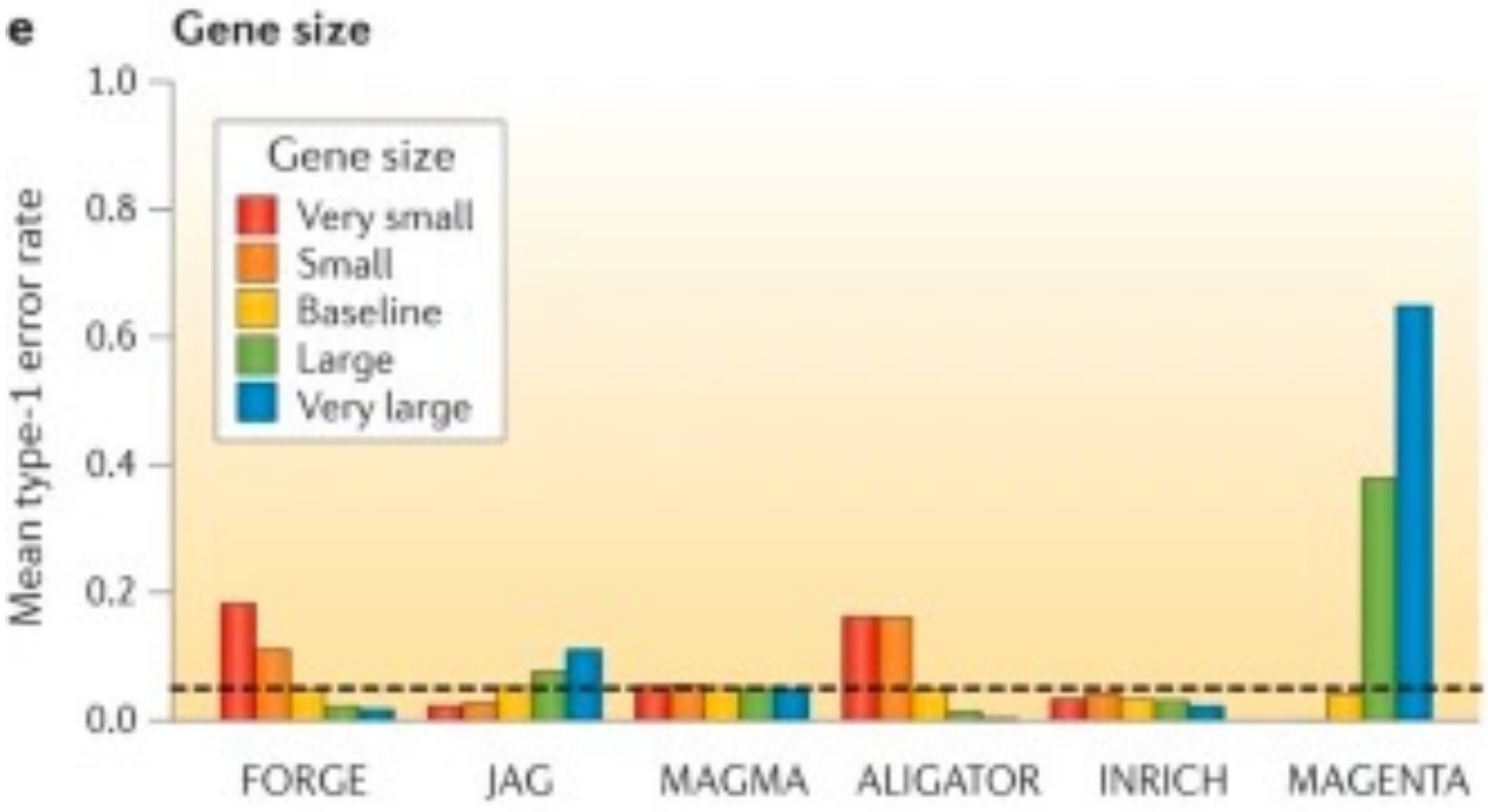


Effect of heritability is dependent on the level of polygenicity - less impact if the heritability is concentrated in a smaller number of effect SNPs

Leeuw et al Nat Rev Gene 2016

Effect of gene size (number of SNPs in gene and amount of SNP LD) on competitive gene-set analysis tools

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

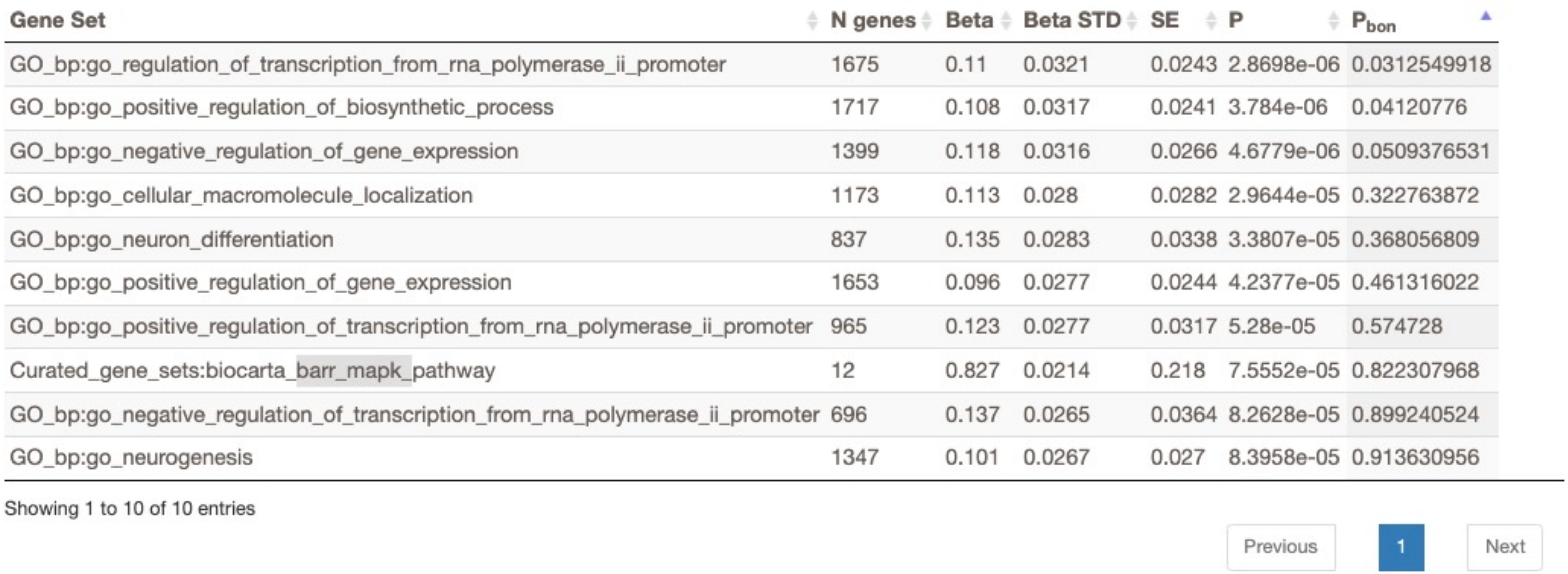


Leeuw et al Nat Rev Gene 2016

MAGMA gene set analysis

- Gene-based *P*-value computed for protein-coding genes by mapping SNPs to genes if they are located within the genes.
- Competitive gene set analysis for 4728 curated gene sets (including canonical pathways) and 6166 GO terms obtained from MsigDB (https://www.gsea-msigdb.org/gsea/msigdb)
- Bonferroni correction (gene) or FDR (gene-set) used to correct for multiple testing.
- 1000G phase 3 is used as a reference panel to calculate LD across SNPs and genes.

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

MAGMA tissue expression analysis

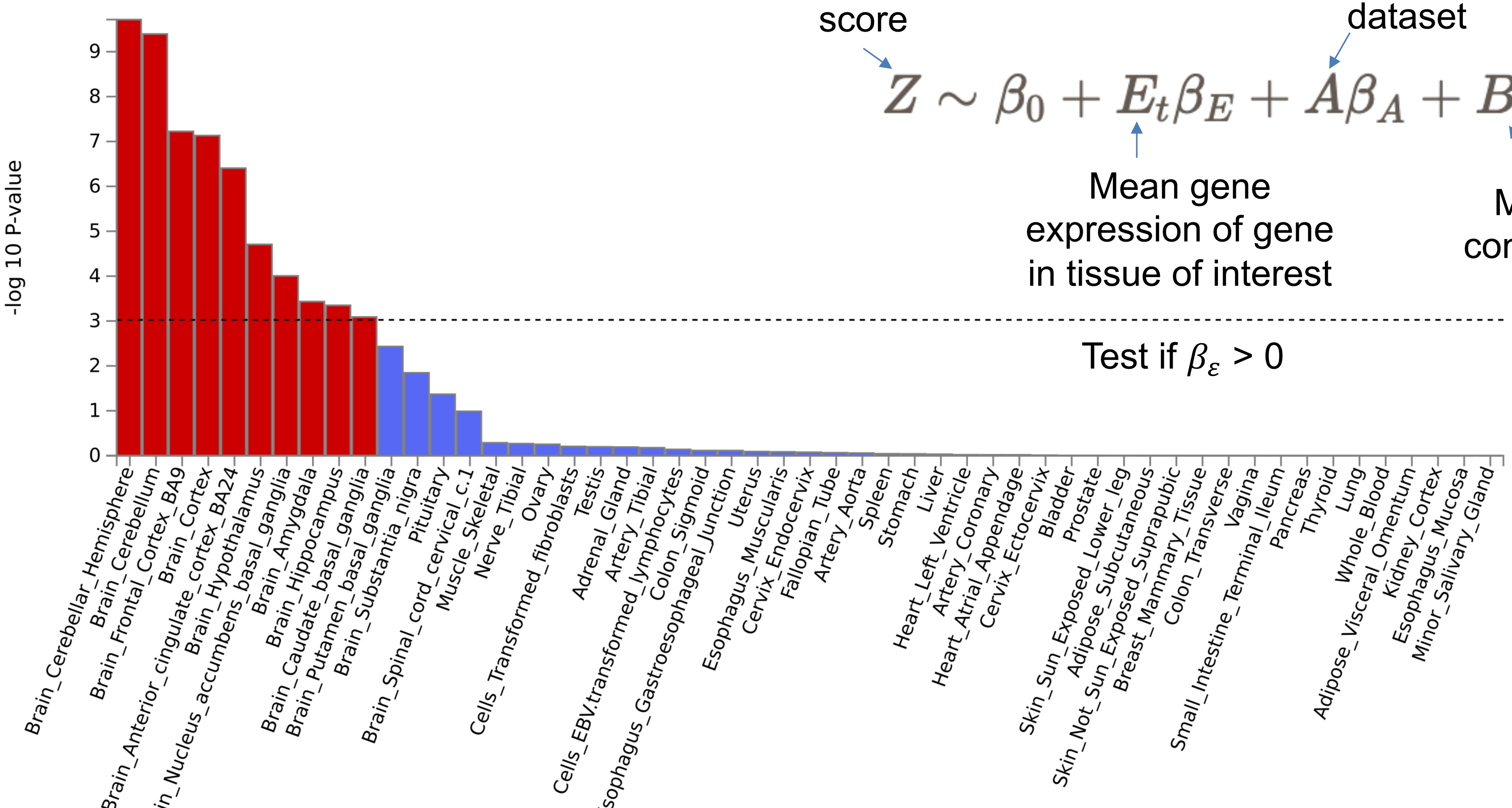
Mean gene expression of gene in all tissues in dataset

Gene-based Z-score

$$Z \sim \beta_0 + E_t\beta_E + A\beta_A + B\beta_B + \epsilon$$

Mean gene expression of gene in tissue of interest

Matrix of confounders



Test if β_ε > 0