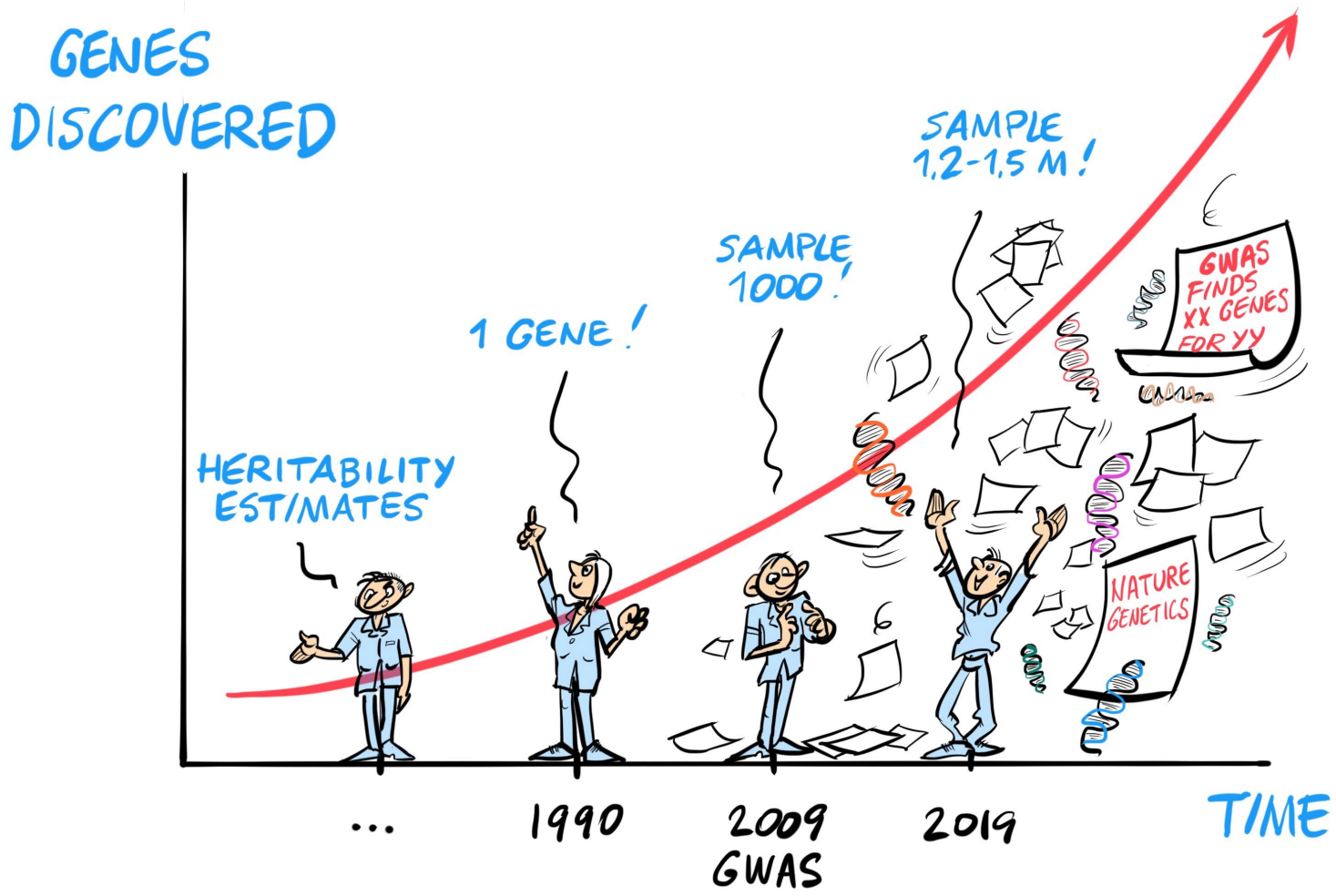
Annotating GWAS results

Danielle Posthuma & Douglas Wightman
CTGlab – VU Amsterdam

shared drive Danielle/FUMA_new2023.ppt

The revolution in genetics

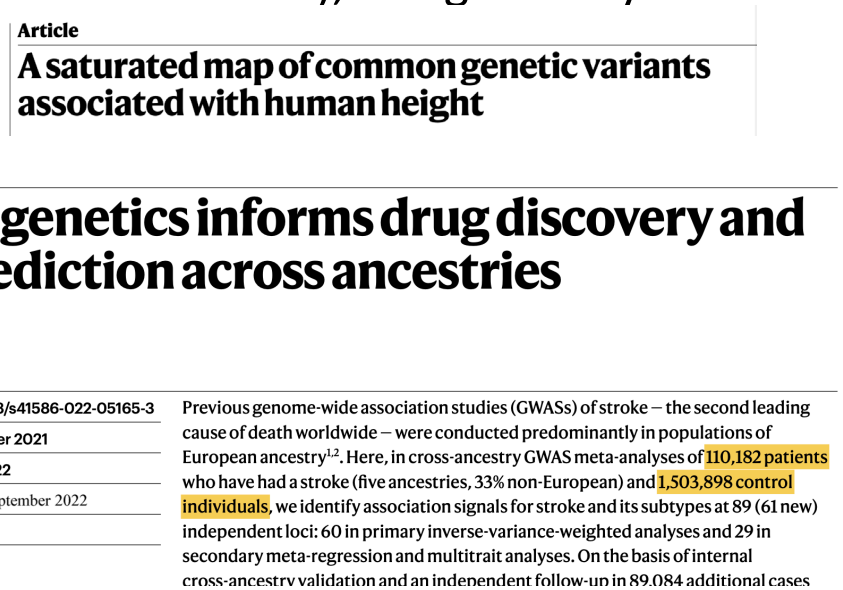


Current genome-wide discovery studies

- They are HUGE: sample sizes typically 100,000's, but even over 1 million is not uncommon!
- Standard quality control and analyses pipelines are thoroughly benchmarked and widely shared
- Good quality GWAS study ALWAYS includes out of sample replication(s) (most journals would not accept without attempted replication)

- There may still be some 'bad quality' GWAS studies (small, replication not included), but generally:

The primary outcome of most GWAS studies is replicable, and highly reliable. Although slight changes in the order of significance may occur with increasing sample size and increasing accuracy, top hits are unlikely to be refuted



Article

A saturated map of common genetic variants associated with human height

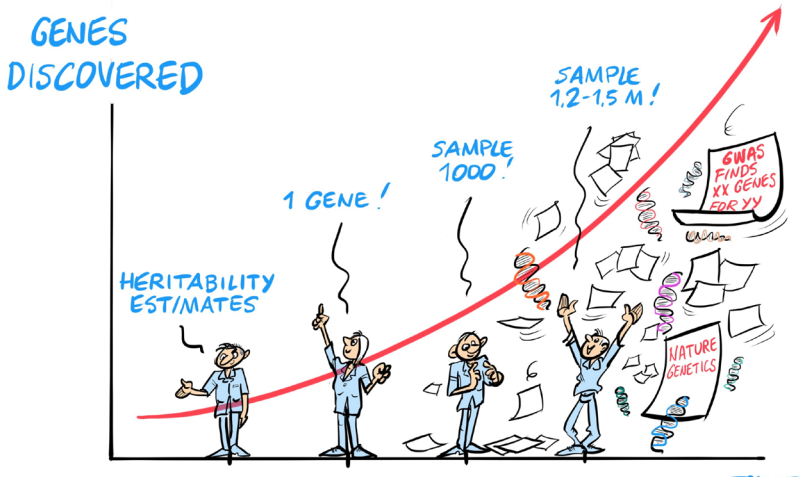
Article

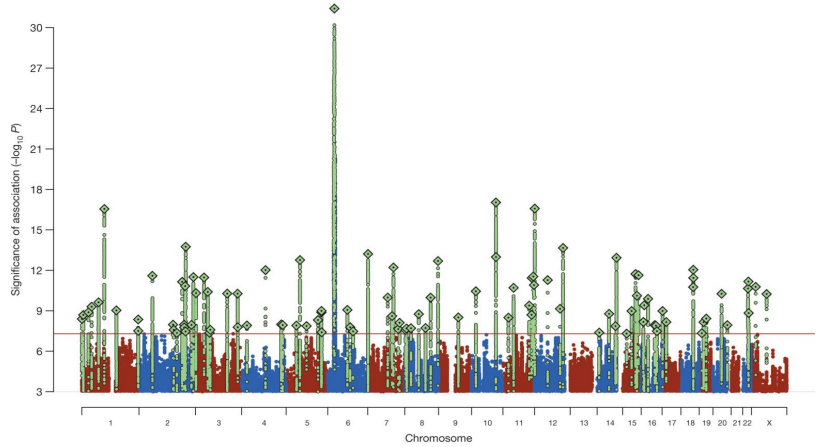
Stroke genetics informs drug discovery and risk prediction across ancestries

https://doi.org/10.1038/s41586-022-05165-3
Received: 15 December 2021
Accepted: 29 July 2022
Published online: 30 September 2022
Open access
Check for updates

Previous genome-wide association studies (GWASs) of stroke – the second leading cause of death worldwide – were conducted predominantly in populations of European ancestry[1,2]. Here, in cross-ancestry GWAS meta-analyses of 110,182 patients who have had a stroke (five ancestries, 33% non-European) and 1,503,898 control individuals, we identify association signals for stroke and its subtypes at 89 (61 new) independent loci: 60 in primary inverse-variance-weighted analyses and 29 in secondary meta-regression and multitrait analyses. On the basis of internal cross-ancestry validation and an independent follow-up in 89,084 additional cases

The revolution in genetics…





S Ripke et al. Nature (2014)

Lots of genes associated with many traits, but..

..unfortunately, this success has not yet translated at the same pace into mechanistic insight.

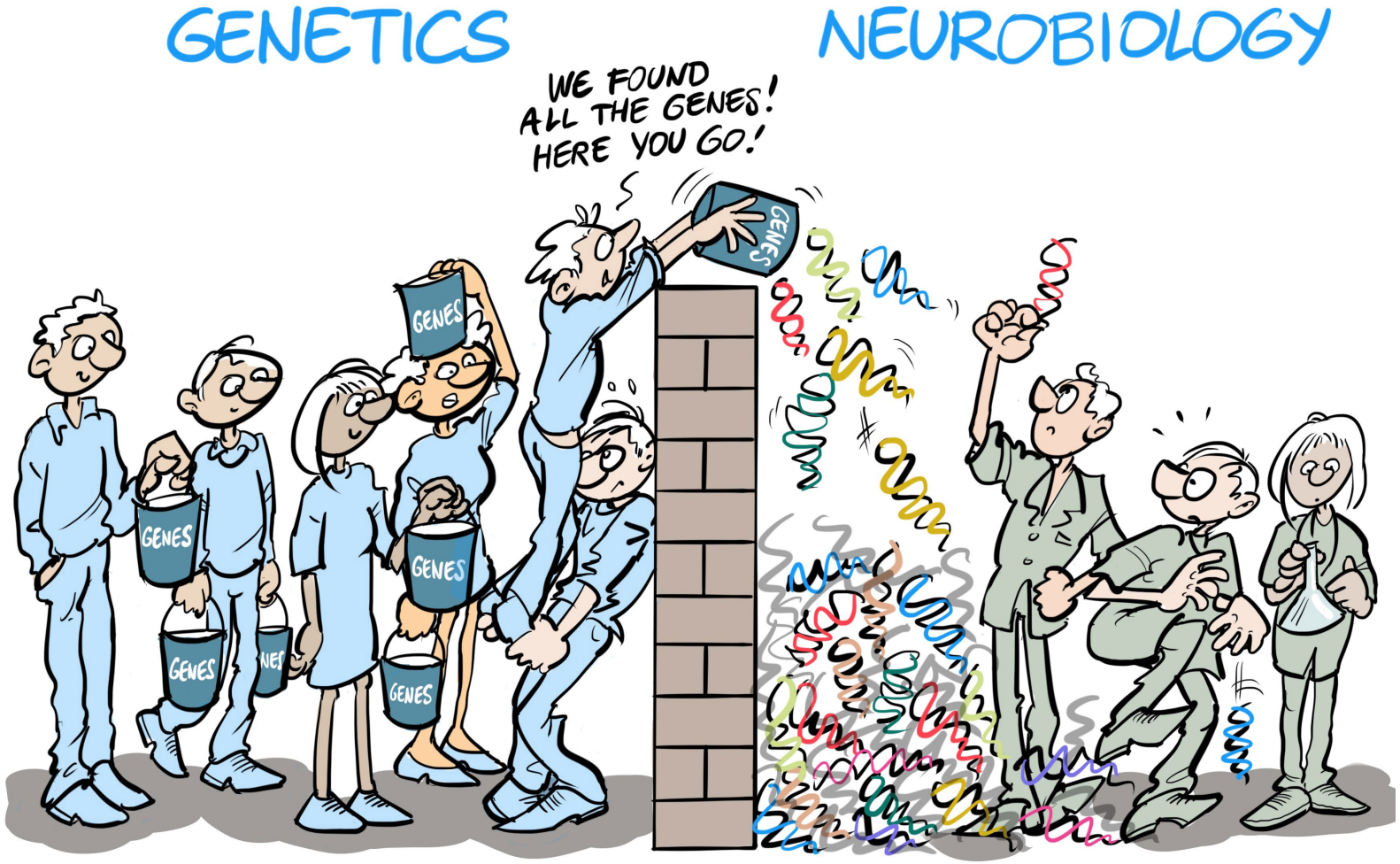
GWAS alone does not provide mechanistic insight: GWAS can provide testable hypotheses on biological mechanisms – functional follow up studies are needed to prove causality/involved mechanisms etc

Why do we need mechanistic insight?

- To understand: if we want to explain why some people fall ill and others don't, we want to know what happens at molecular, cellular and organismal levels
- To treat: mechanistic insight may aid in developing novel treatments by providing starting points

For prediction we do not necessarily need mechanistic insight though.
But - prediction makes most sense if it can be tailored to treatment

Why is translation of GWAS into mechanistic insight so hard?



Challenges in interpreting GWAS outcome

1. *Correlation between variants (LD)*

Significant association P-values are distributed over **blocks of correlated genetic variants**: actual causal variant is unclear

Challenges in interpreting GWAS outcome

1. *Correlation between variants (LD)*

Significant association P-values are distributed over **blocks of correlated genetic variants**: actual causal variant is unclear

Solution: - **functional annotation** (functional variants more likely to be causal than non-functional ones, e.g. tools FUMA, VEP, ANNOVAR)

- **statistical fine-mapping** (the known correlation structure can be modeled against the observed pattern of association values to pinpoint the most likely causal SNPs, this can be integrated with functional information (e.g. tools FINEMAP, PAINTOR)

Challenges in interpreting GWAS outcome

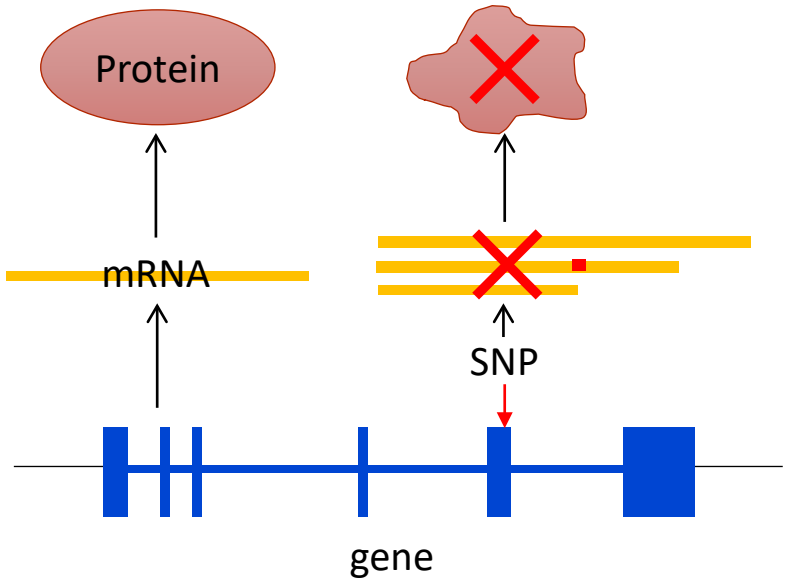
2. *Many GWAS hits are in* ***non-coding regions***

The majority of GWAS hits are not in exonic regions but are intronic or non-genic. That means they do not directly lead to a different protein structure and their impact on protein function may be less straightforward to assess.

Coding and Non-coding SNPs

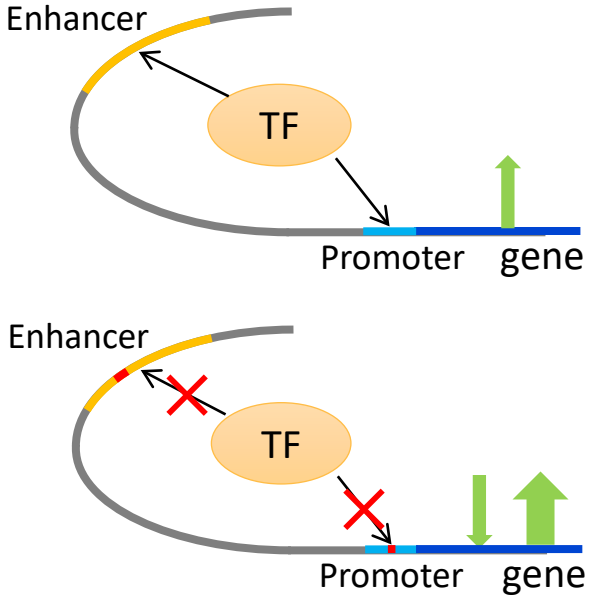
Coding SNPs
Non-synonymous SNPs
Loss of Function (Stop gain, Stop loss)
Splicing SNPs

Direct effect on gene product/proteins



Non-coding SNPs
Enhancer/Promoter regions
Transcription Factor Binding Sites
(TFBSs)

Indirect effect on gene products



Expression quantitative loci (eQTLs)

Challenges in interpreting GWAS outcome

2. *Many GWAS hits are in **non-coding regions***

The majority of GWAS hits are not in exonic regions but are intronic or non-genic. That means they do not directly lead to a different protein structure and their impact on protein function may be less straightforward to assess.

Solution: link GWAS variants to genes via **functional annotation** using regulatory information from external resources, such as GTEX (e-QTL), chromatin interactions, i.e. add information on the association of a variant with DNA transcription and RNA or protein levels

Challenges in interpreting GWAS outcome

3. *Many traits are **polygenic***

When a trait is polygenic, multiple genetic variants of small effect contribute. A single genetic variant, even if it is known to be causal, is usually not informative for biology

Challenges in interpreting GWAS outcome

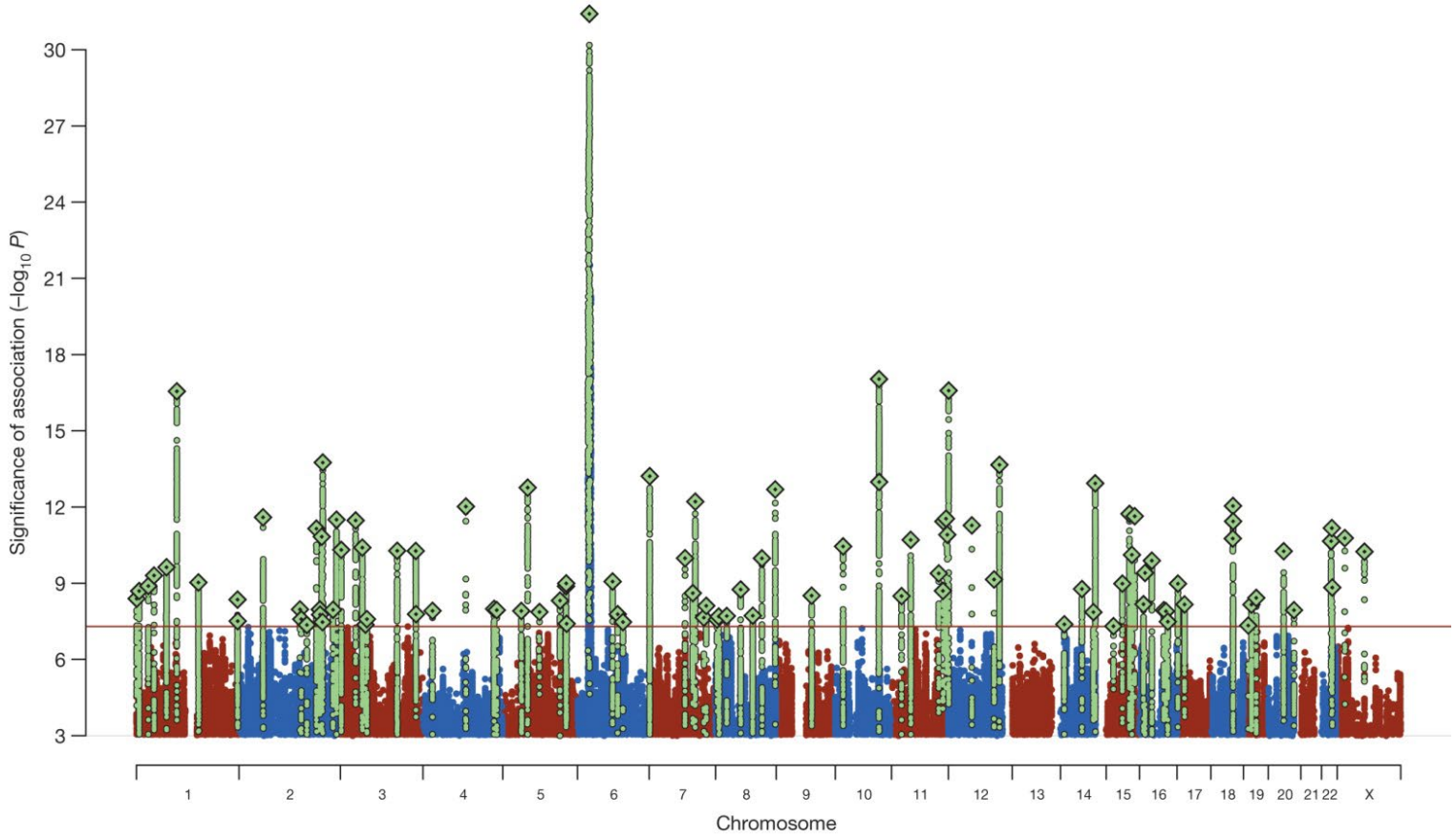
3. *Many traits are **polygenic***

When a trait is polygenic, multiple genetic variants of small effect contribute. A single genetic variant, even if it is known to be causal, is usually not informative for biology

Solution: map associated SNPs to genes and look for **convergence** in biological pathways, shared cellular or synaptic function, co-localization, co-expression in tissue or cell types (e.g. tools MAGMA, Ldscore regression, DEPICT)

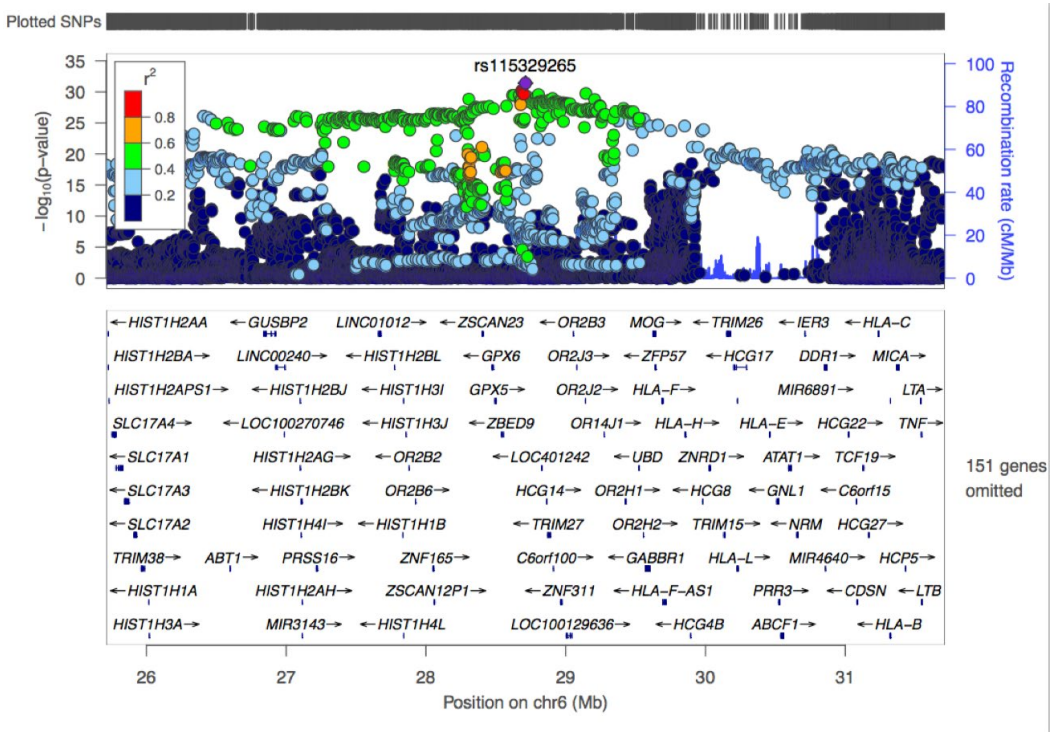
This will be covered in next session on gene-set analysis

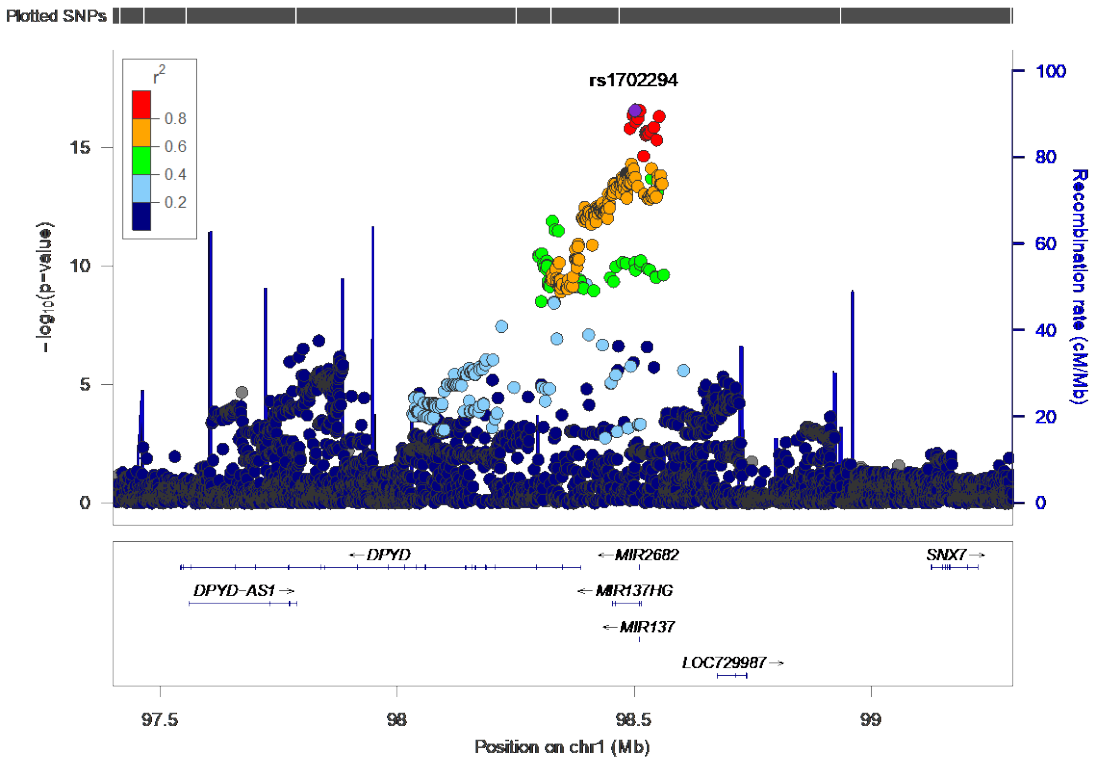
Awesome GWAS results



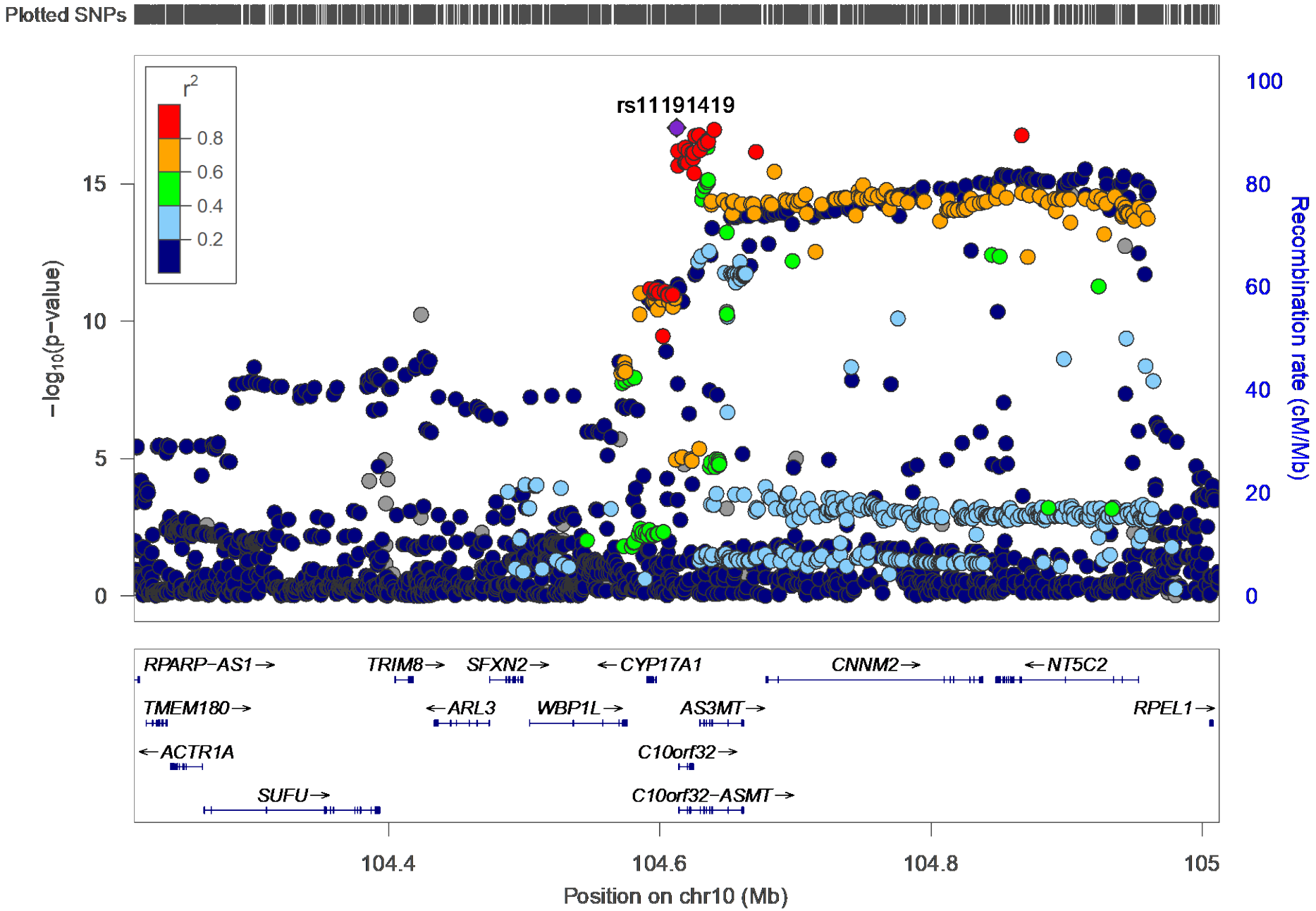
S Ripke *et al. Nature* (2014)

Awesome GWAS results - but how to interpret?

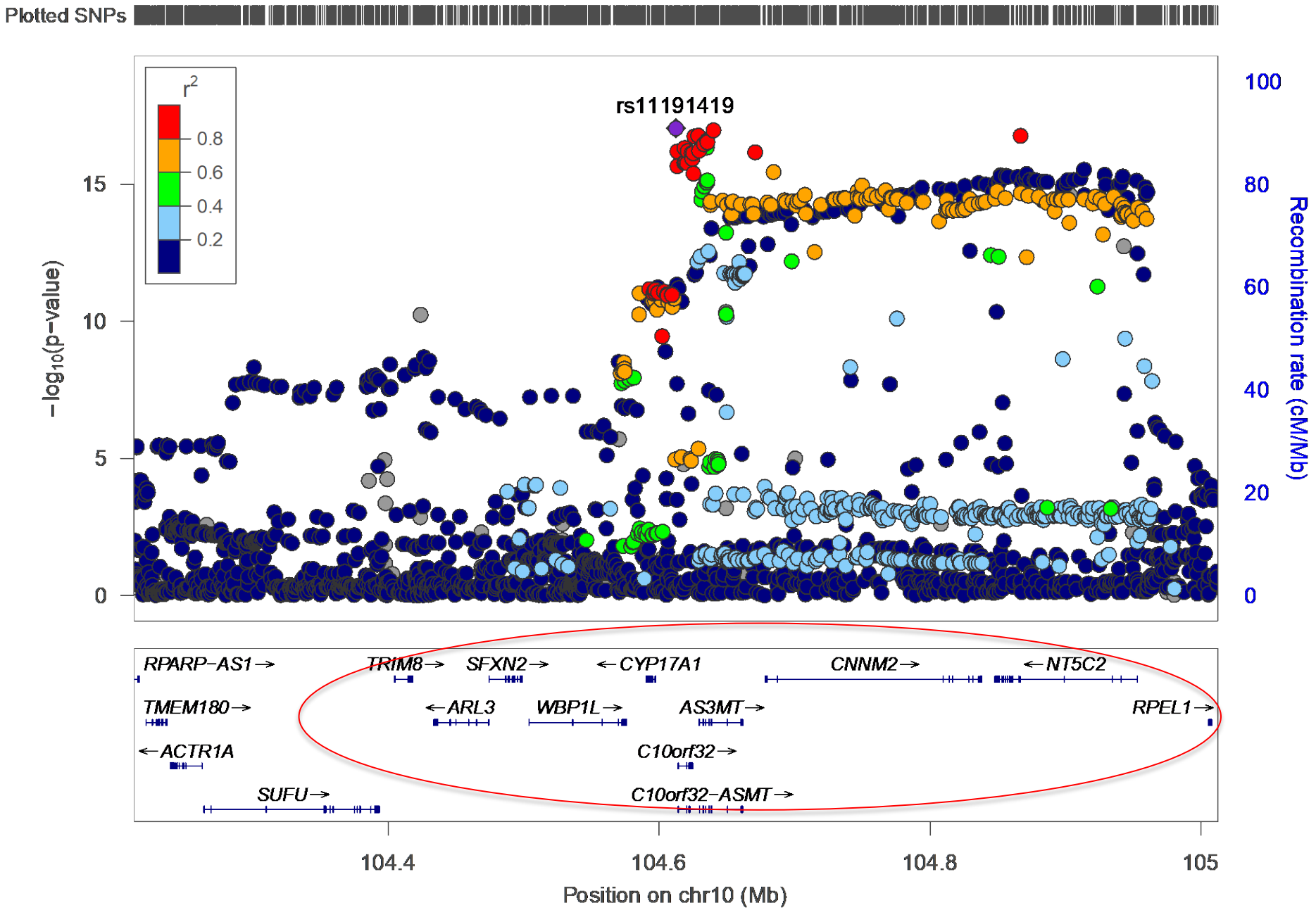




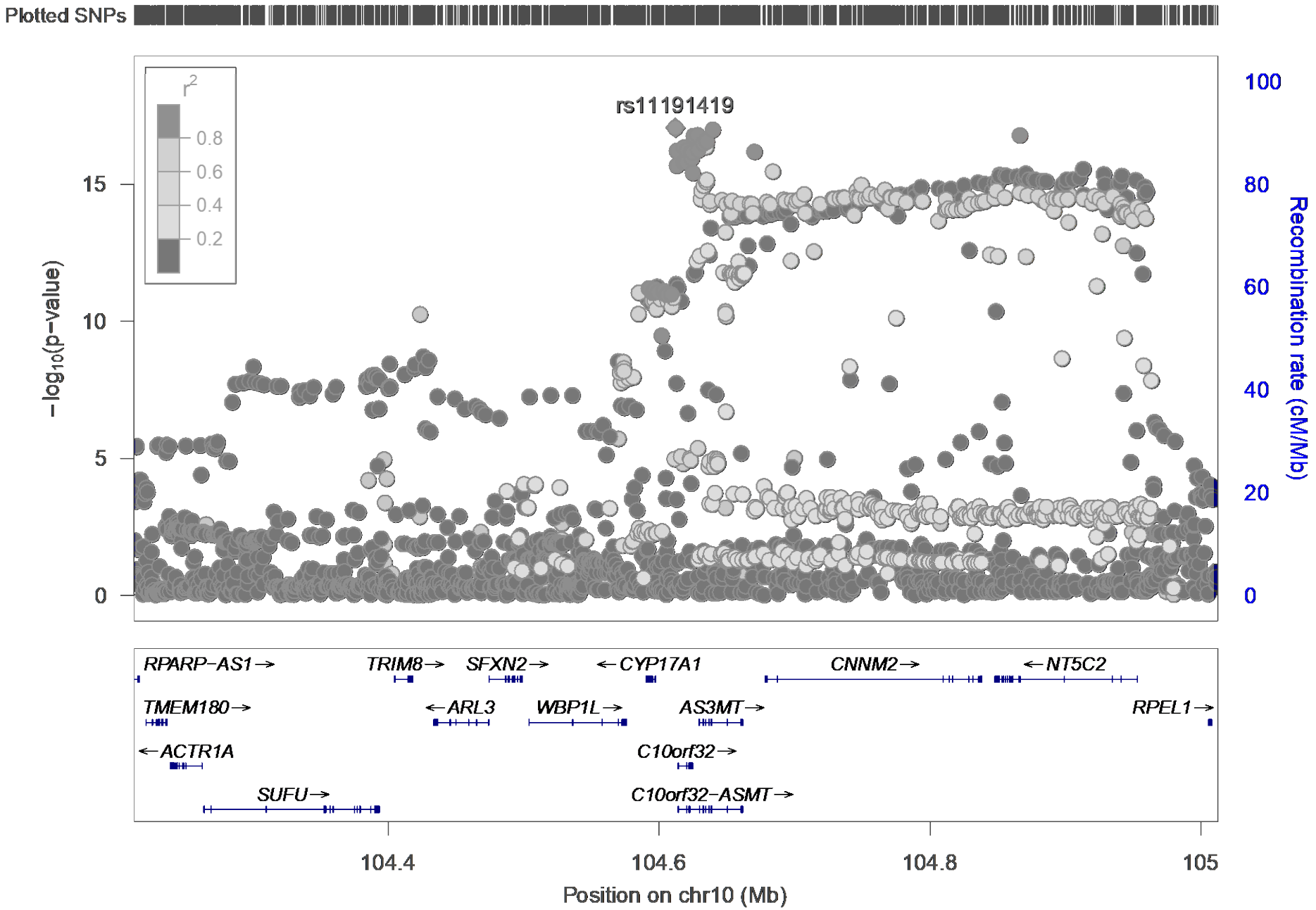
GWAS risk locus



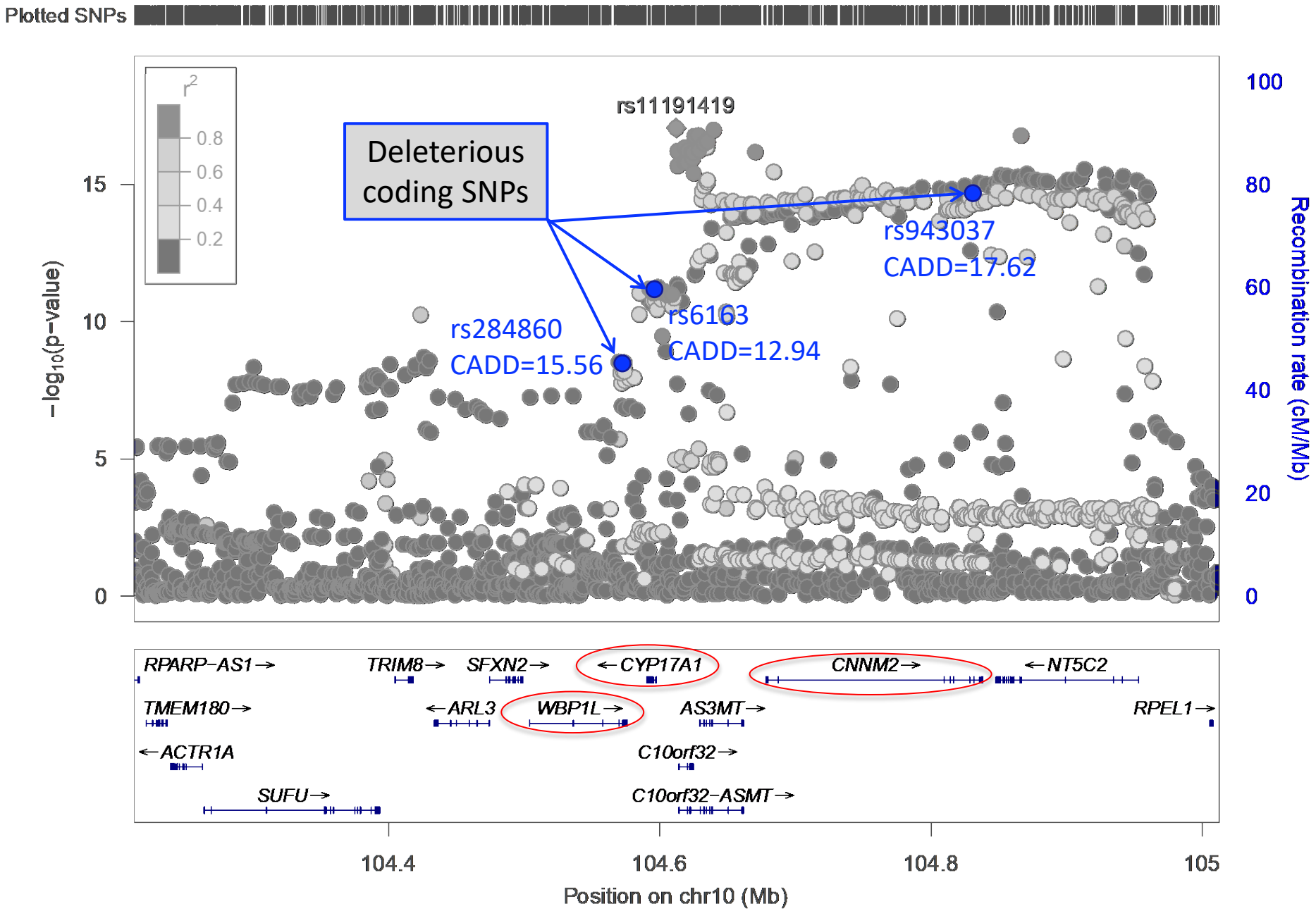
GWAS risk locus



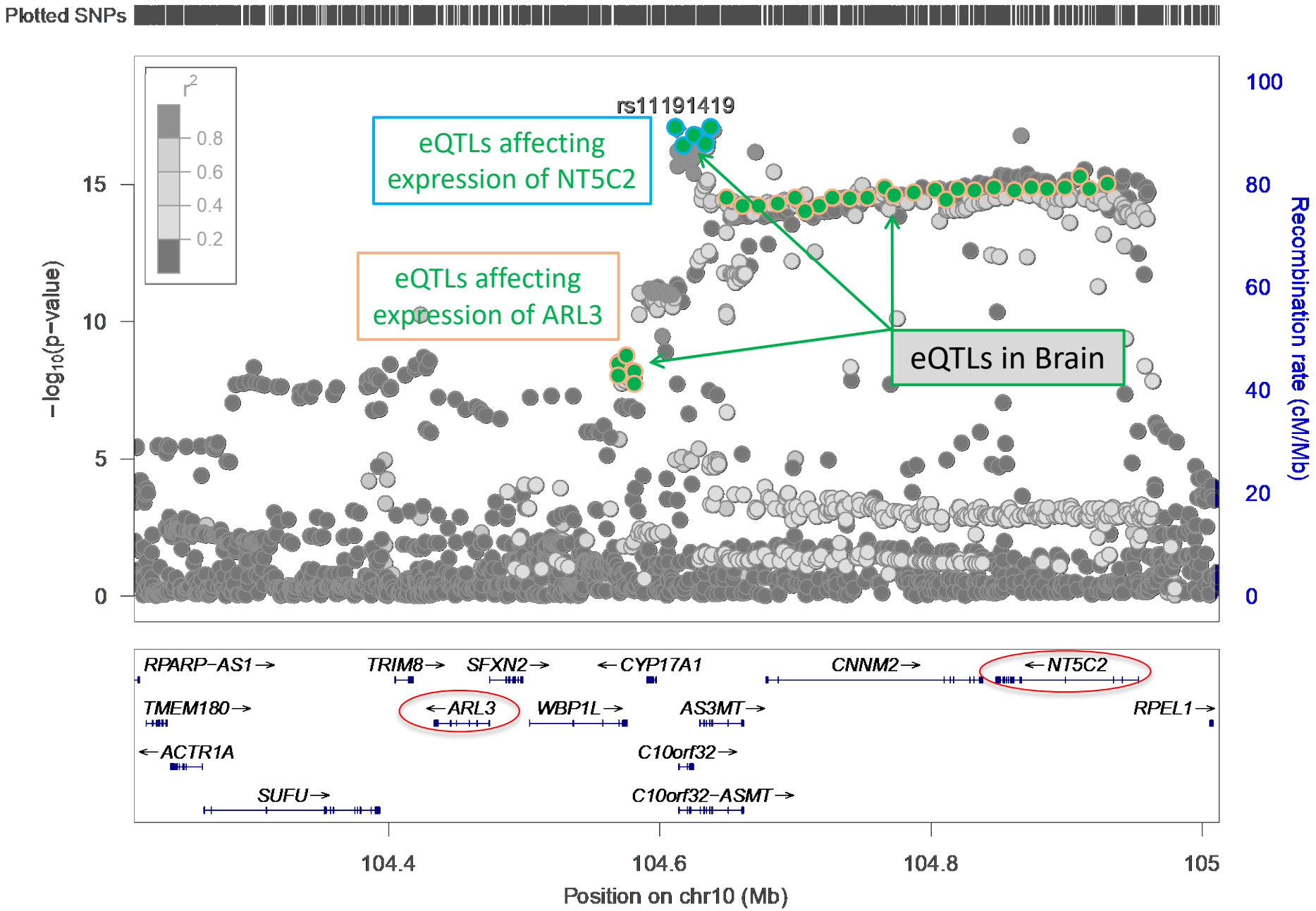
GWAS risk locus



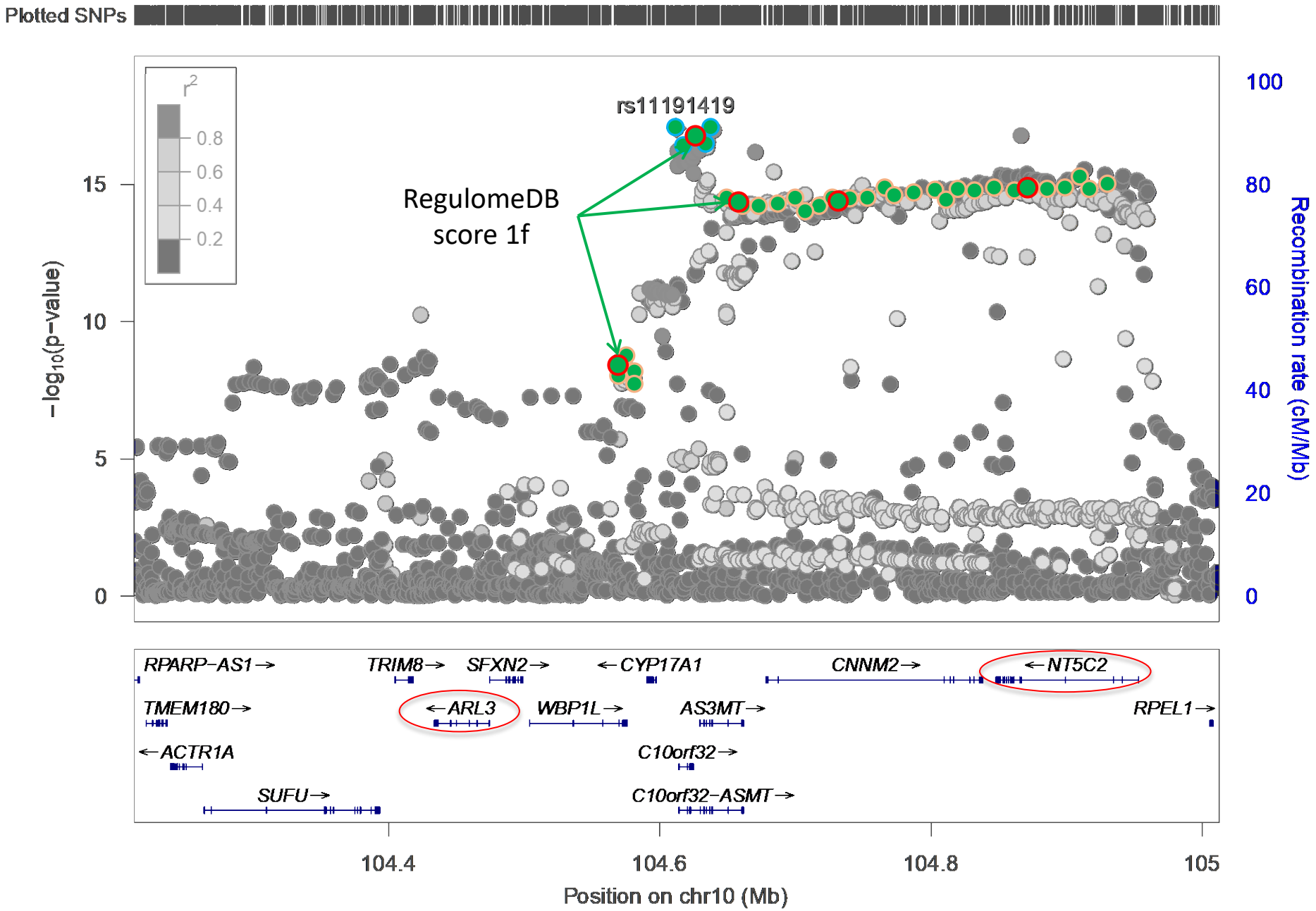
GWAS risk locus



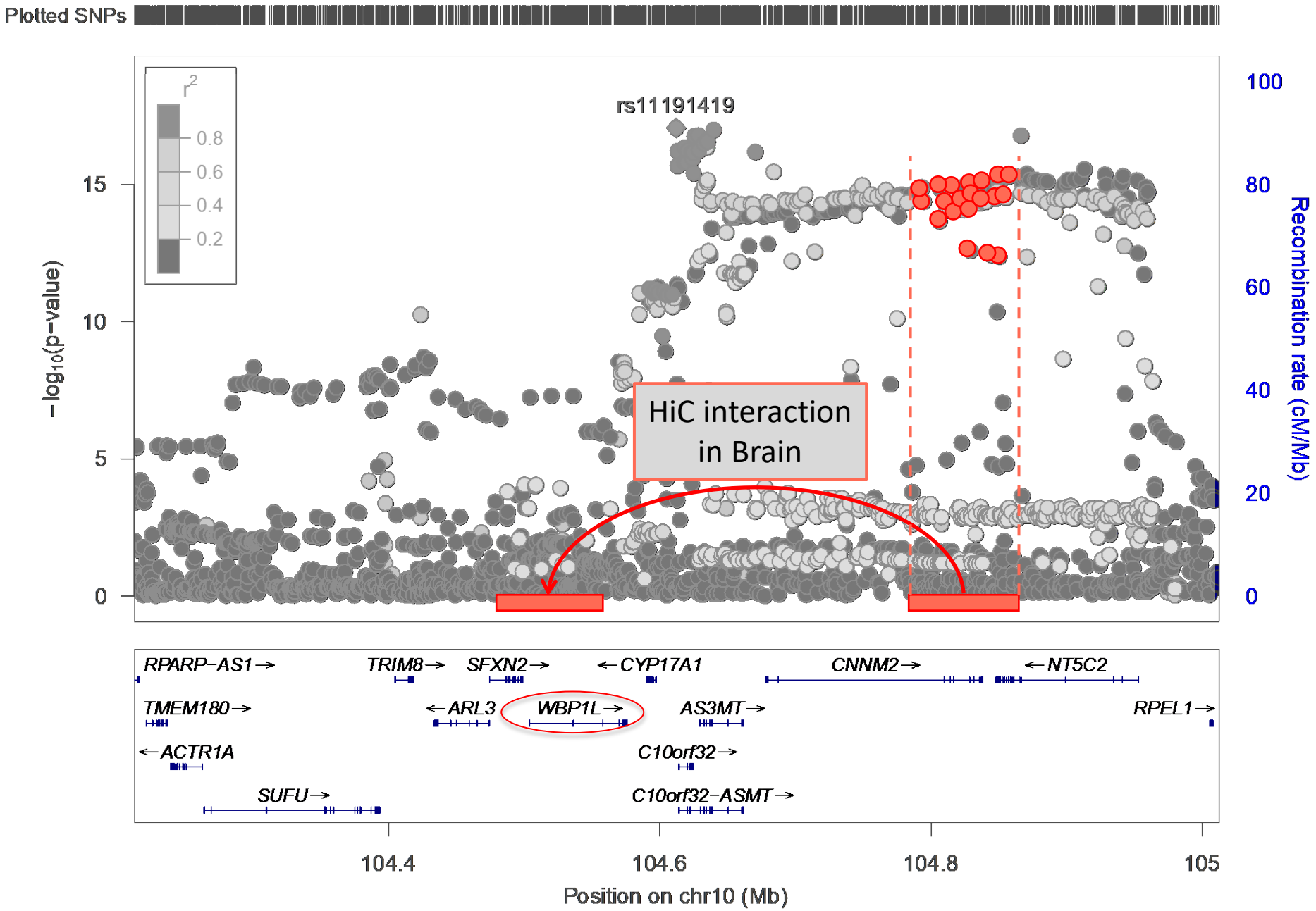
GWAS risk locus



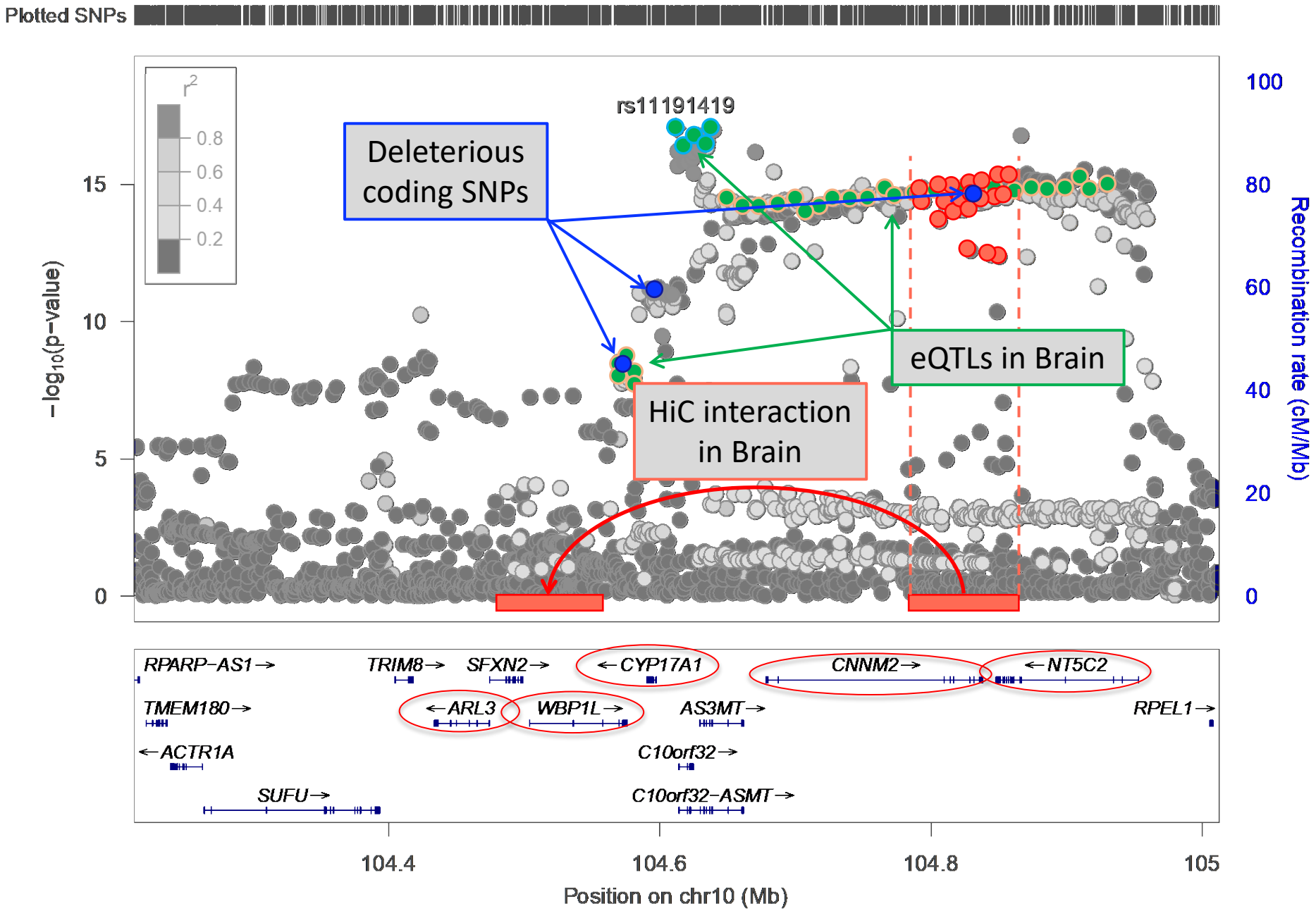
GWAS risk locus



GWAS risk locus



GWAS risk locus



Post-GWAS annotations and prioritization

PLINK

1. **Linkage disequilibrium (LD)**
   Identify all SNPs in LD with significant hits.

Post-GWAS annotations and prioritization

PLINK

1. **Linkage disequilibrium (LD)**
   Identify all SNPs in LD with significant hits.

2. **Variant annotation**
   Functional consequence on genes (e.g. exonic, intronic or splicing site)

ANNOVAR

Post-GWAS annotations and prioritization

PLINK

1. **Linkage disequilibrium (LD)**
   Identify all SNPs in LD with significant hits.

2. **Variant annotation**
   Functional consequence on genes e.g. exonic, intronic or splicing site)

ANNOVAR

3. **Functional annotation**
   Deleteriousness, regulatory elements and epigenetic data

CADD

GTEx Portal

BBMRI.nl
Biobanking and BioMolecular resources Research Infrastructure The Netherlands

ENCODE

RegulomeDB

ROADMAP epigenomics PROJECT

HiC

Post-GWAS annotations and prioritization

PLINK

1. **Linkage disequilibrium (LD)**
   Identify all SNPs in LD with significant hits.

2. **Variant annotation**
   Fu

3. **F**
   I

4. **F**
   Tiss

Multiple databases
Multiple software
Multiple steps
Reformatting of data

Time-consuming + error prone

MSigDB
Molecular Signatures Database

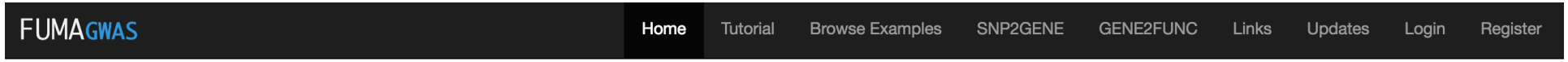
WIKIPATHWAYS
Pathways for the People

KEGG
Kyoto Encyclopedia of Genes and Genomes

GEO
Gene Expression Omnibus

FUMA: Functional Mapping and Annotation of genetic associations

Available at http://fuma.ctglab.nl



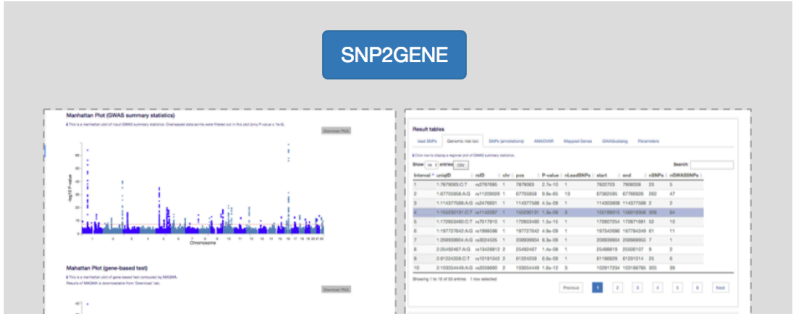
FUMA GWAS

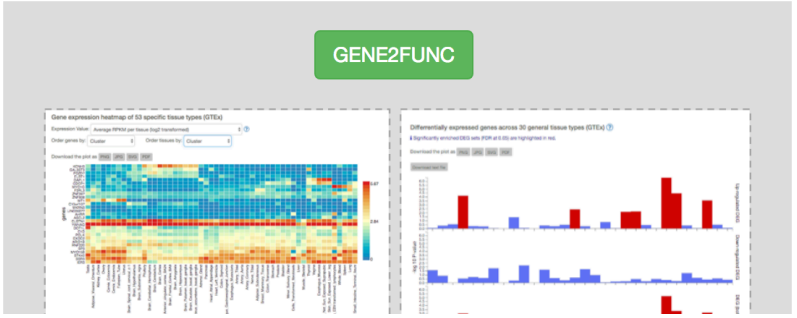
Functional Mapping and Annotation of Genome-Wide Association Studies

FUMA is a platform that can be used to annotate, prioritize, visualize and interpret GWAS results.
The SNP2GENE function takes GWAS summary statistics as an input, and provides extensive functional annotation for all SNPs in genomic areas identified by lead SNPs.
The GENE2FUNC function takes a list of gene IDs (as identified by SNP2GENE or as provided manually) and annotates genes in biological context
To submit your own GWAS, login is required for security reason. If you have't registered yet, you can do from here.
You can browse example results of FUMA for a few GWAS from Browse Examples without registration or login.

Please post any questions, suggestions and bug reports on Google Forum: FUMA GWAS users.

Citation:
When using FUMA, please cite the following.
K. Watanabe, E. Taskesen, A. van Bochoven and D. Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**:1826. (2017).
https://www.nature.com/articles/s41467-017-01261-5
Depending on which results you are going to report, please also cite the original study of data sources/tools used in FUMA (references are available at Links).





Developed by Kyoko Watanabe

FUMA

Upload GWAS summary statistics

Use default settings or adjust as desired

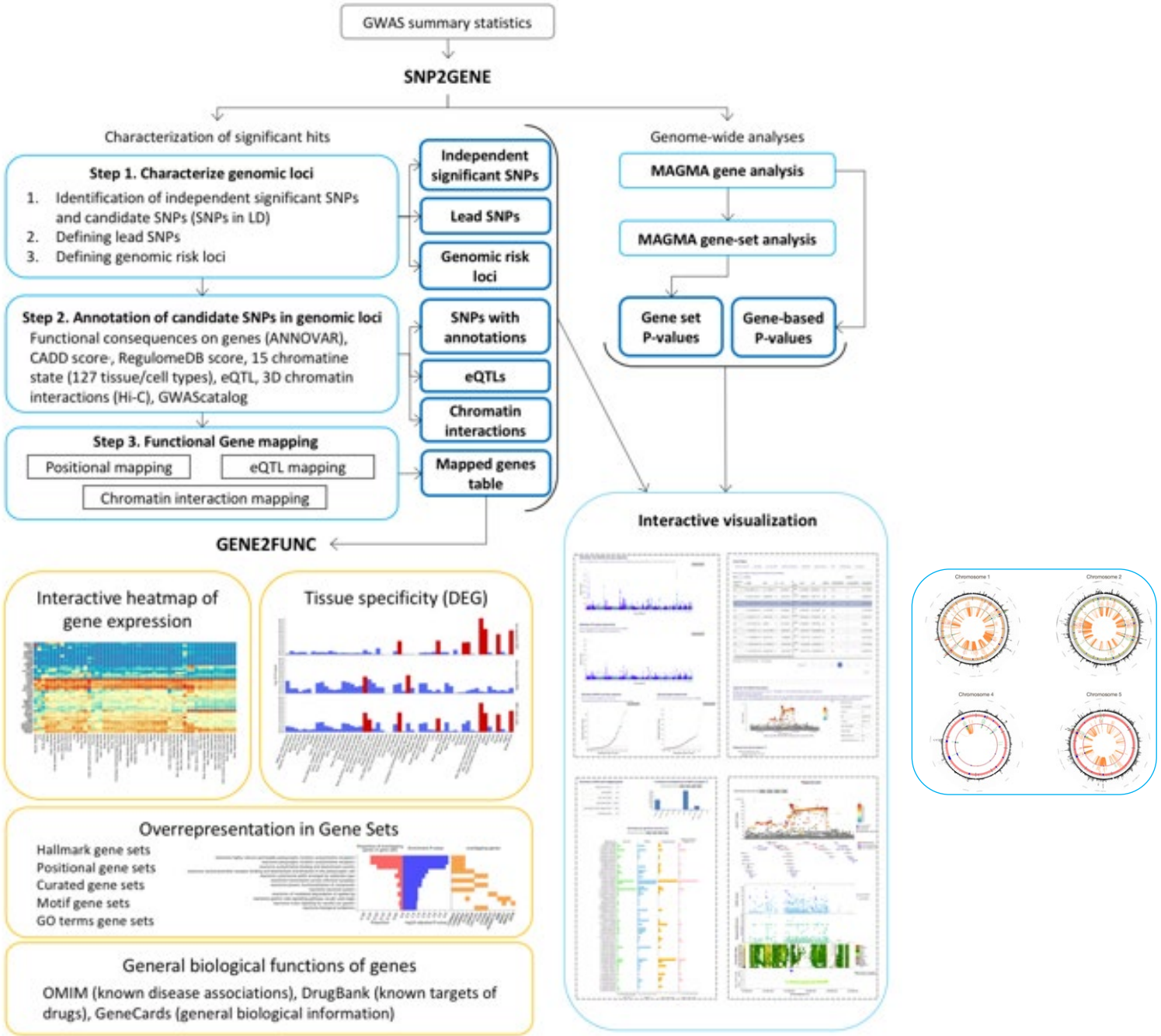
Output:

- Defined risk loci
- Manhattan & QQ plots
- Gene-based and gene-set analyses, incl tissue and cell type
- Extensive annotations of all SNPs in risk loci in LD with top SNPs
- Variant to gene mapping using position, eQTL, chromatin interaction
- Look-up of SNPs in gwas catalogue
- Lots of visualizations and downloadable tables

Upload list of gene names

Output:

- gene-set and tissue enrichments



FUMA

Upload GWAS summary statistics

Use default settings or adjust as desired

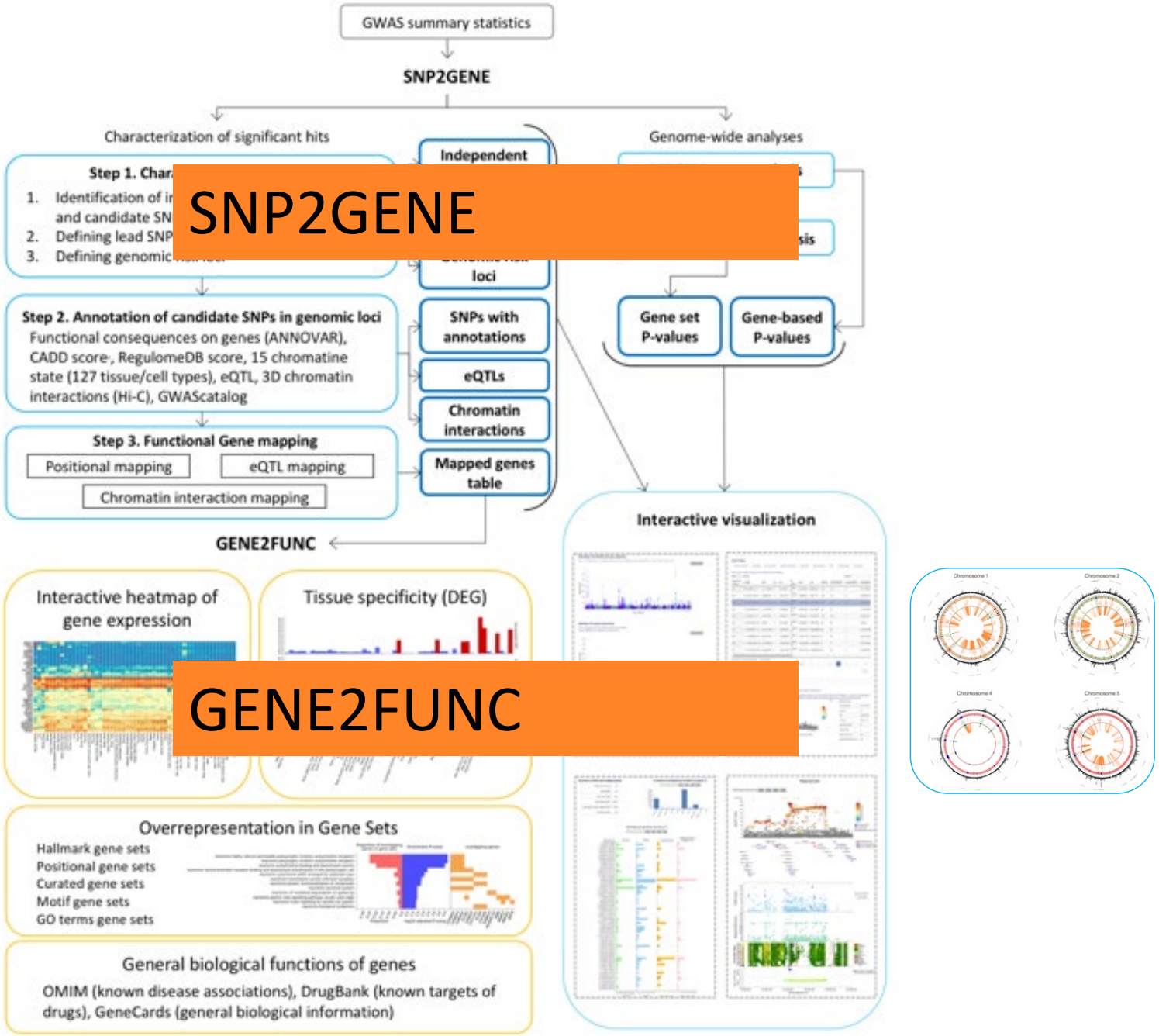
Output:

- Defined risk loci
- Manhattan & QQ plots
- Gene-based and gene-set analyses, incl tissue and cell type
- Extensive annotations of all SNPs in risk loci in LD with top SNPs
- Variant to gene mapping using position, eQTL, chromatin interaction
- Look-up of SNPs in gwas catalogue
- Lots of visualizations and downloadable tables

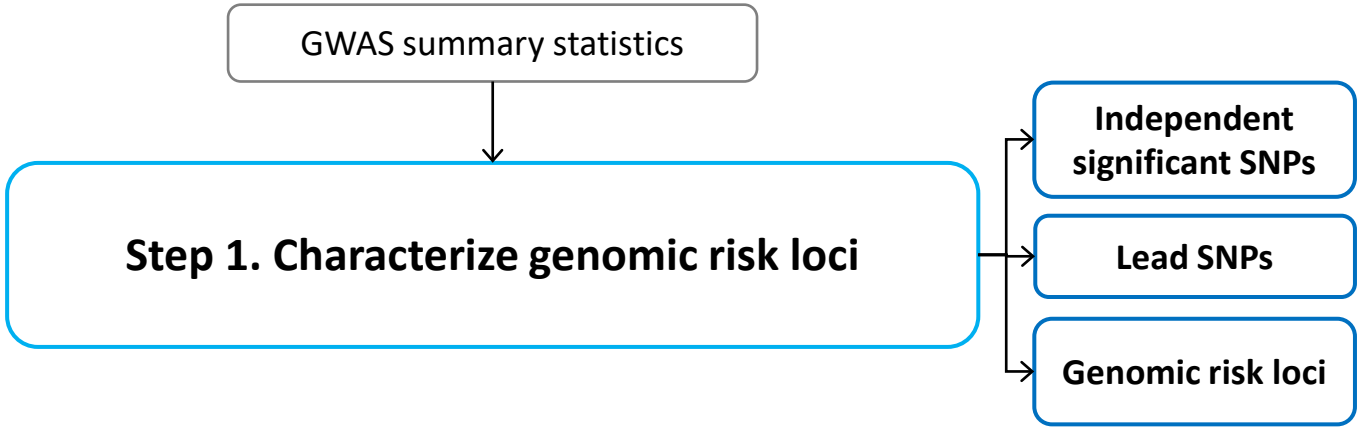
Upload list of gene names

Output:

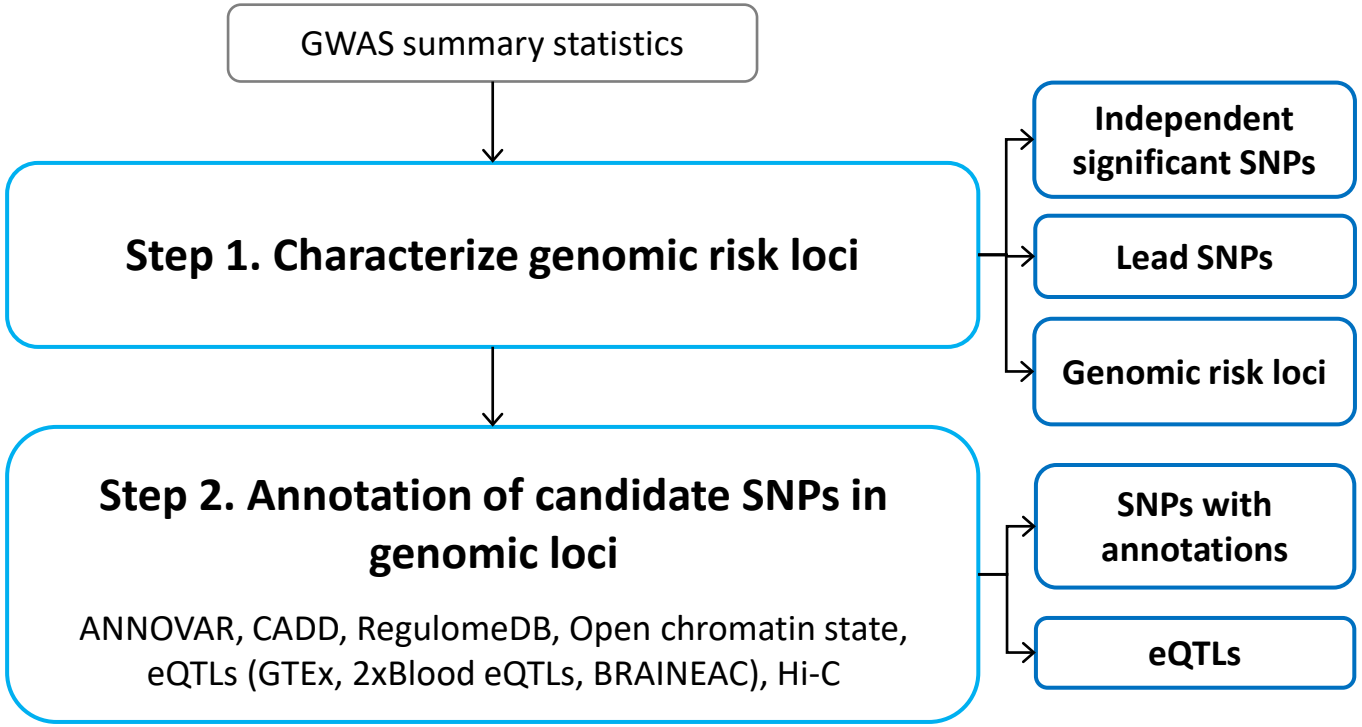
- gene-set and tissue enrichments



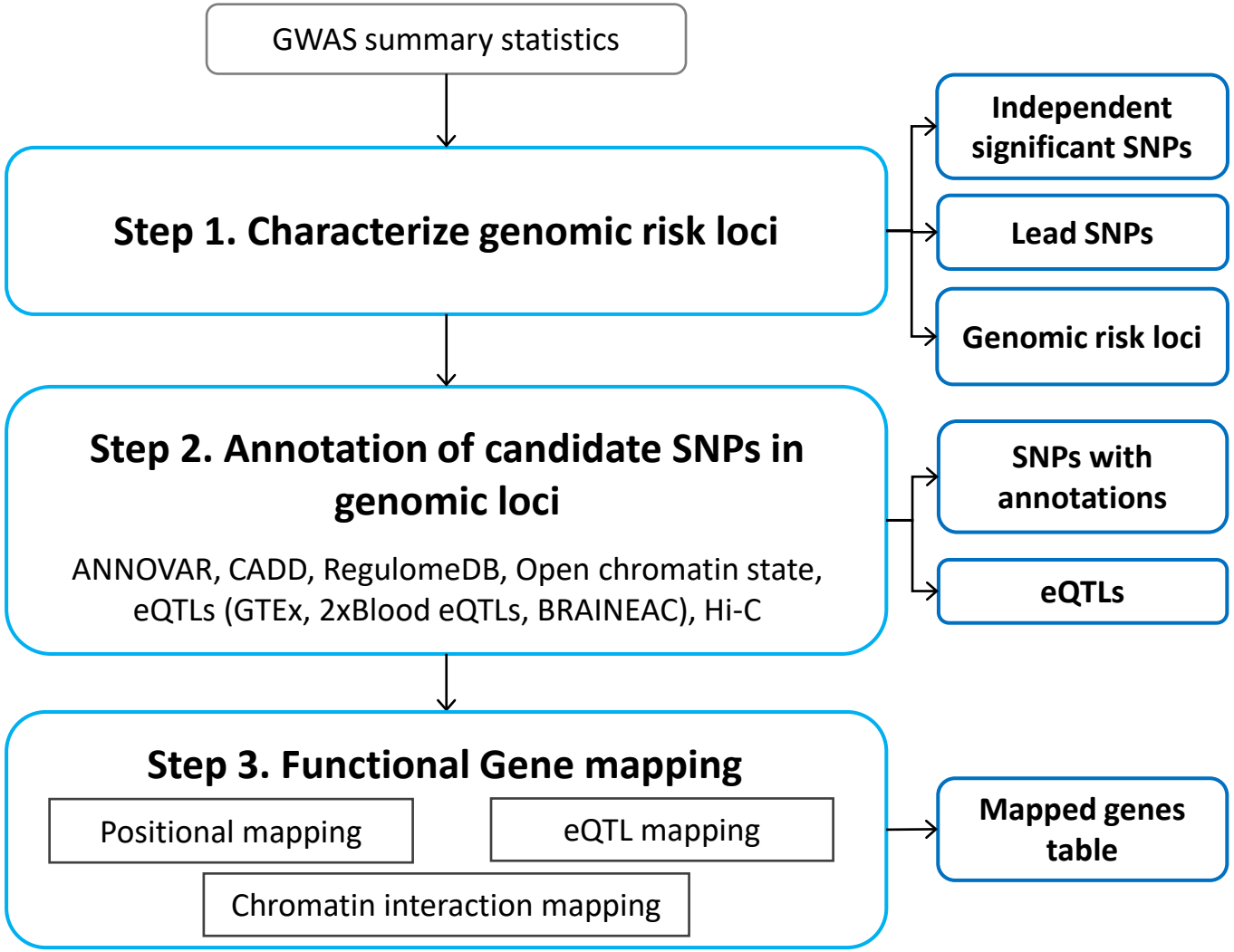
SNP2GENE


GWAS summary statistics
Step 1. Characterize genomic risk loci
Independent significant SNPs
Lead SNPs
Genomic risk loci

SNP2GENE

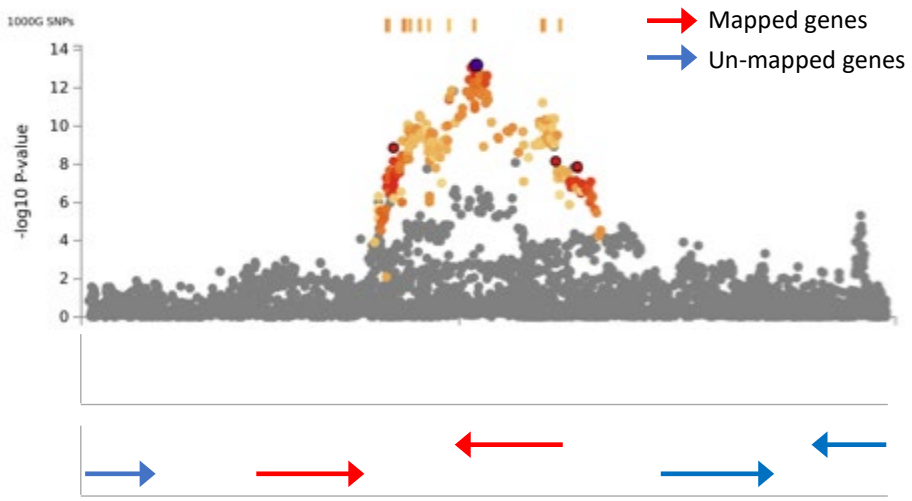

GWAS summary statistics
Step 1. Characterize genomic risk loci
Independent significant SNPs
Lead SNPs
Genomic risk loci
Step 2. Annotation of candidate SNPs in genomic loci
ANNOVAR, CADD, RegulomeDB, Open chromatin state, eQTLs (GTEx, 2xBlood eQTLs, BRAINEAC), Hi-C
SNPs with annotations
eQTLs

SNP2GENE


GWAS summary statistics
Step 1. Characterize genomic risk loci
Independent significant SNPs
Lead SNPs
Genomic risk loci
Step 2. Annotation of candidate SNPs in genomic loci
ANNOVAR, CADD, RegulomeDB, Open chromatin state, eQTLs (GTEx, 2xBlood eQTLs, BRAINEAC), Hi-C
SNPs with annotations
eQTLs
Step 3. Functional Gene mapping
Positional mapping
eQTL mapping
Chromatin interaction mapping
Mapped genes table

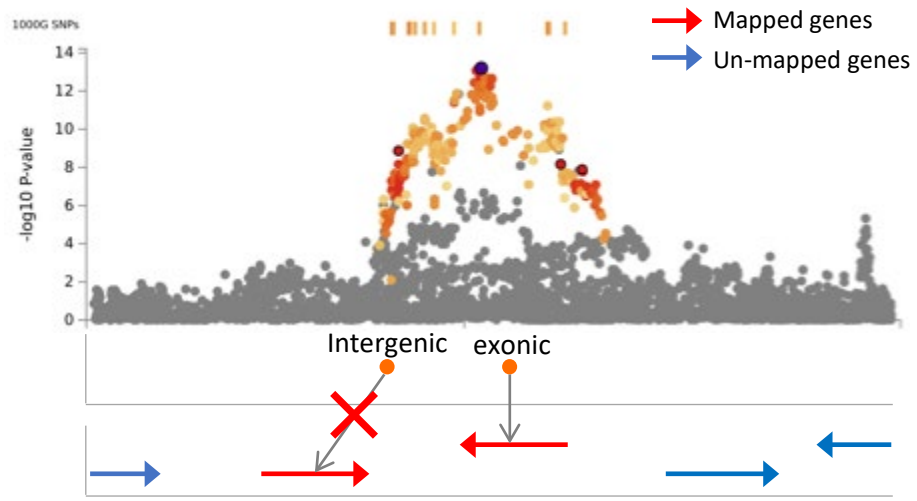
Gene Mapping

Positional mapping



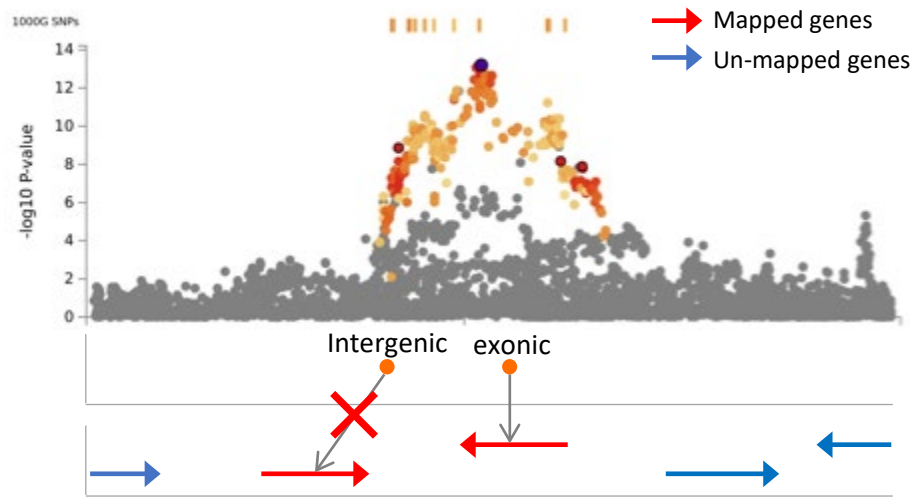
Gene Mapping

Positional mapping

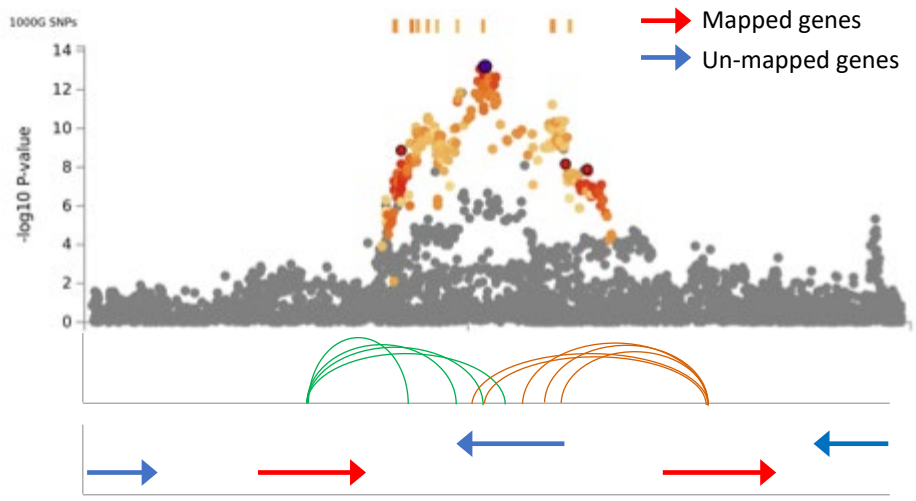


Gene Mapping

Positional mapping

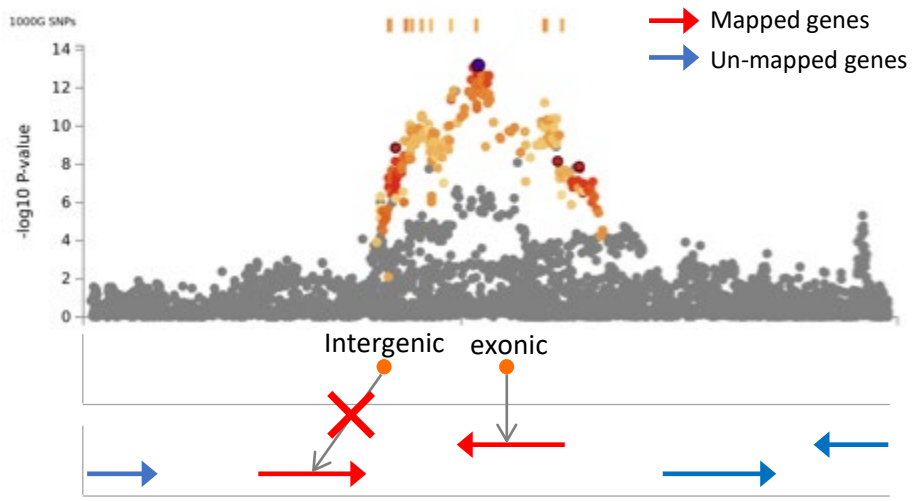


eQTL mapping

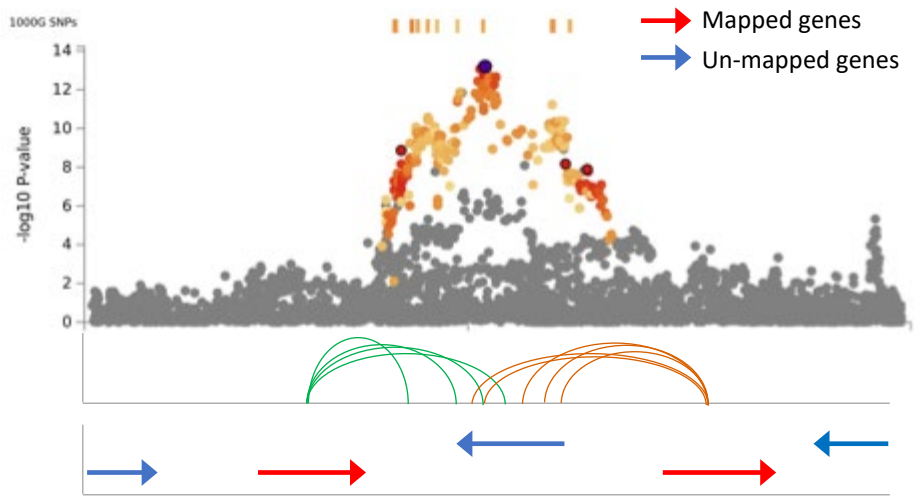


Gene Mapping

Positional mapping



eQTL mapping



Optional filtering of SNPs based on functional annotations prior to gene mapping

CADD score (>12.37 is considered as highly deleterious)

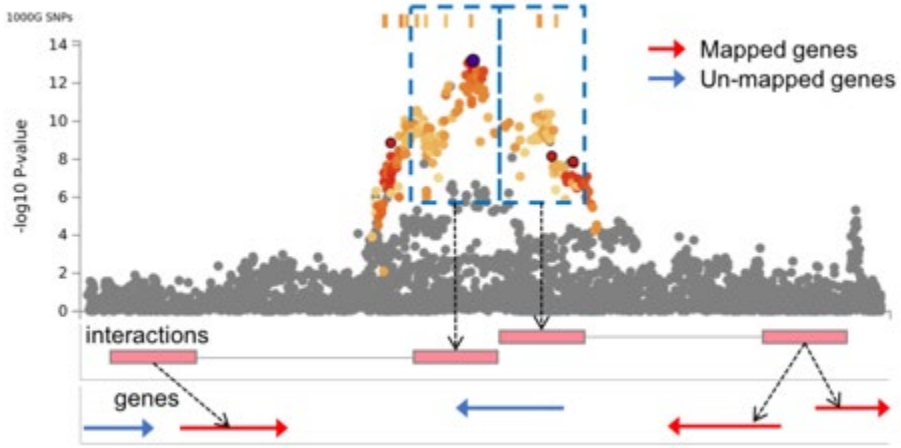
RegulomeDB score (categorical score from 1a to 7)

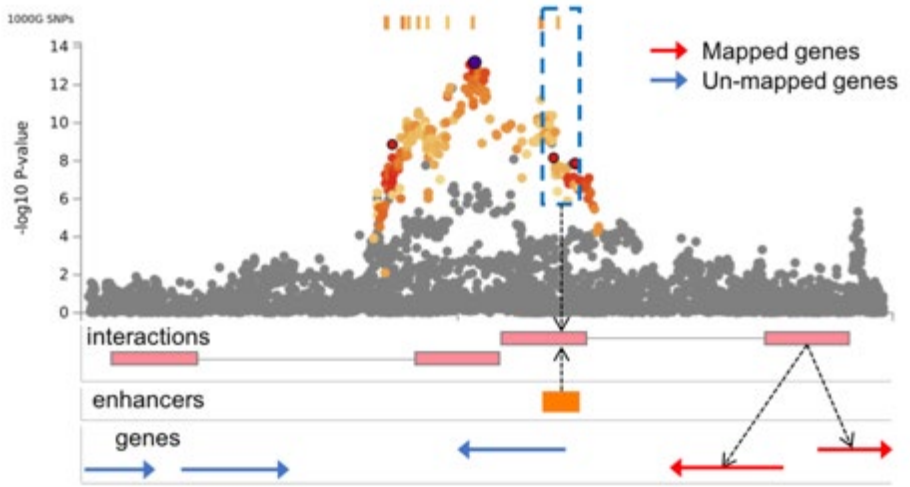
15-core chromatin status (1-7 states are considered as open, tissue specific)

*Parameters can be set for each mapping separately

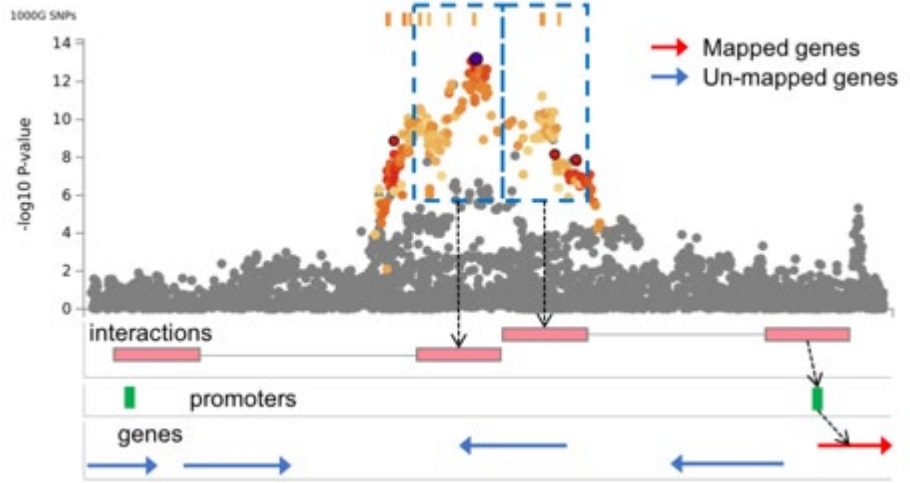
Gene Mapping

Chromatin interaction mapping

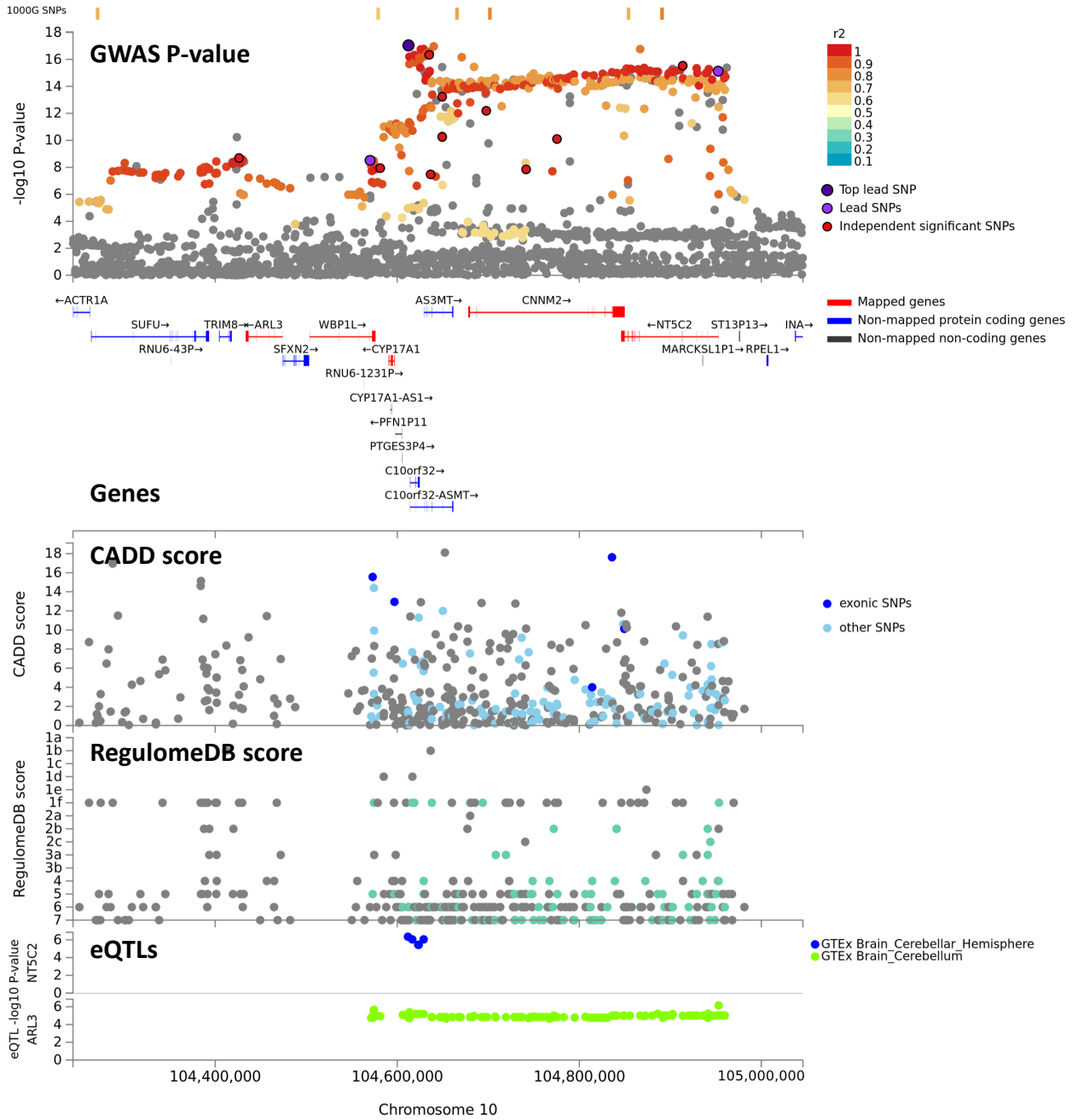




Map SNPs in a genomic region interacting with promoter region of genes (250bp up- and 500bp downstream of TSS) in selected tissue types.
SNPs can be further filtered on overlap with predicted enhancers or promotor regions from Roadmap.

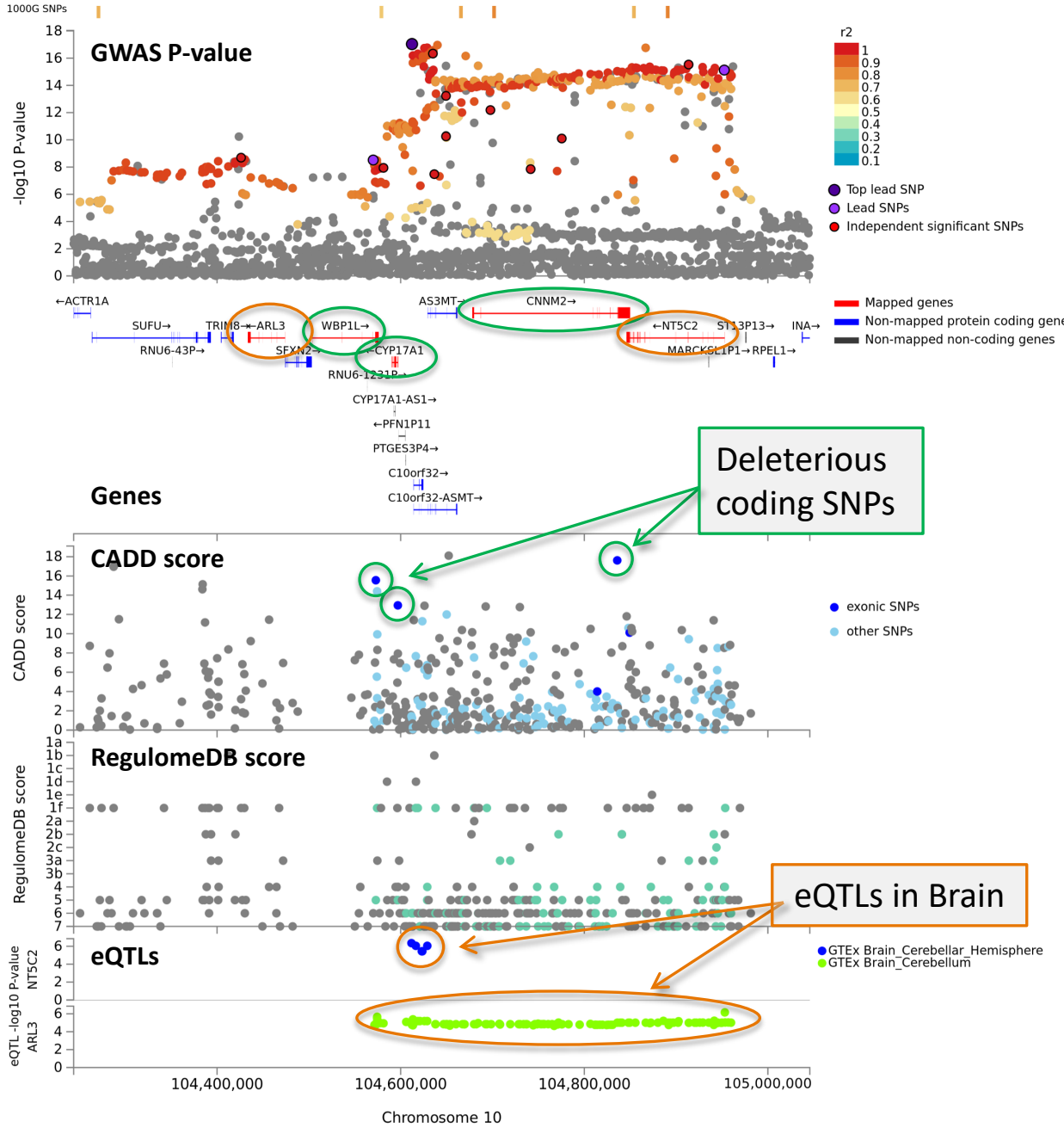


FUMA: regional plot

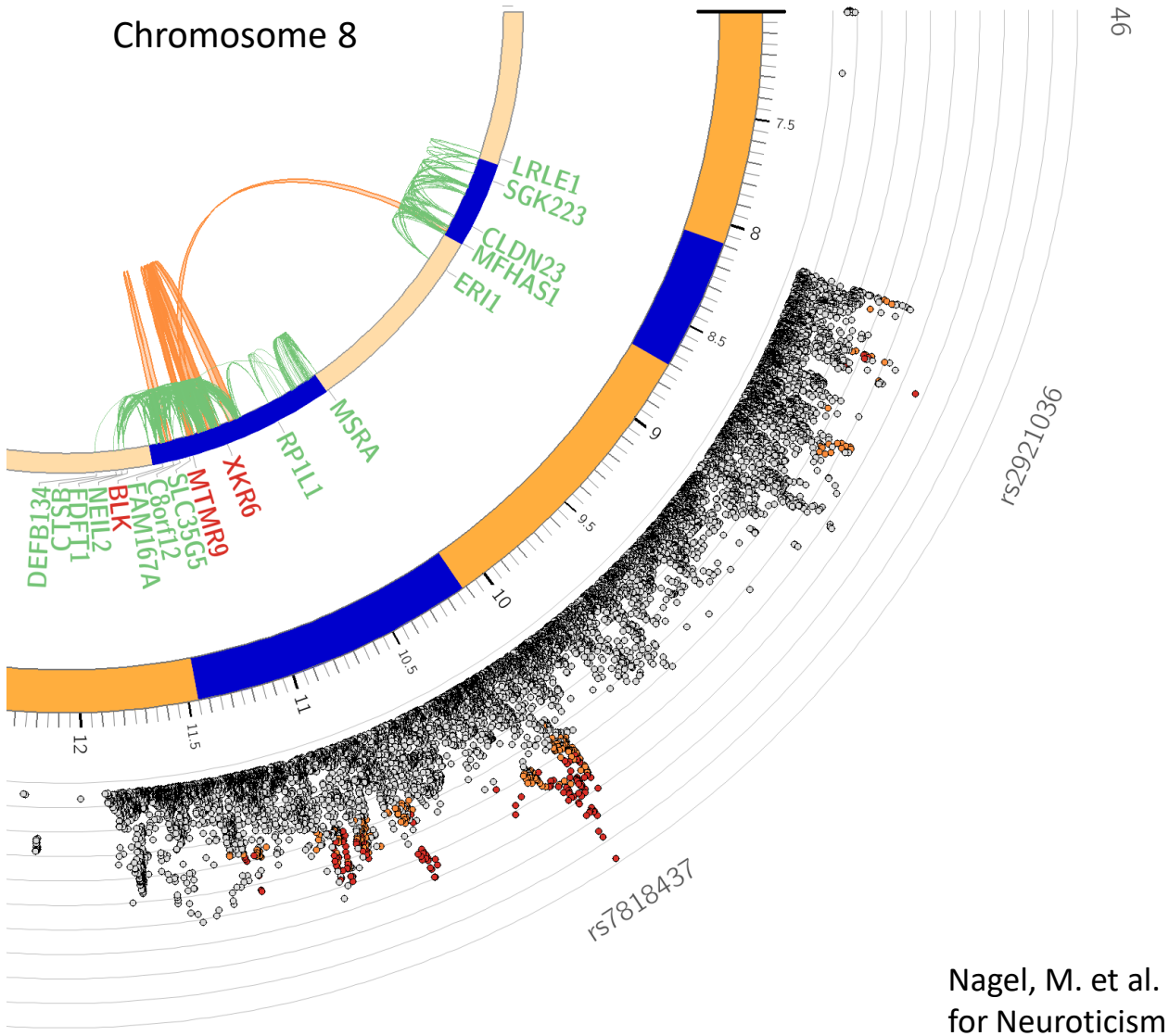


FUMA: regional plot

FUMA: regional plot



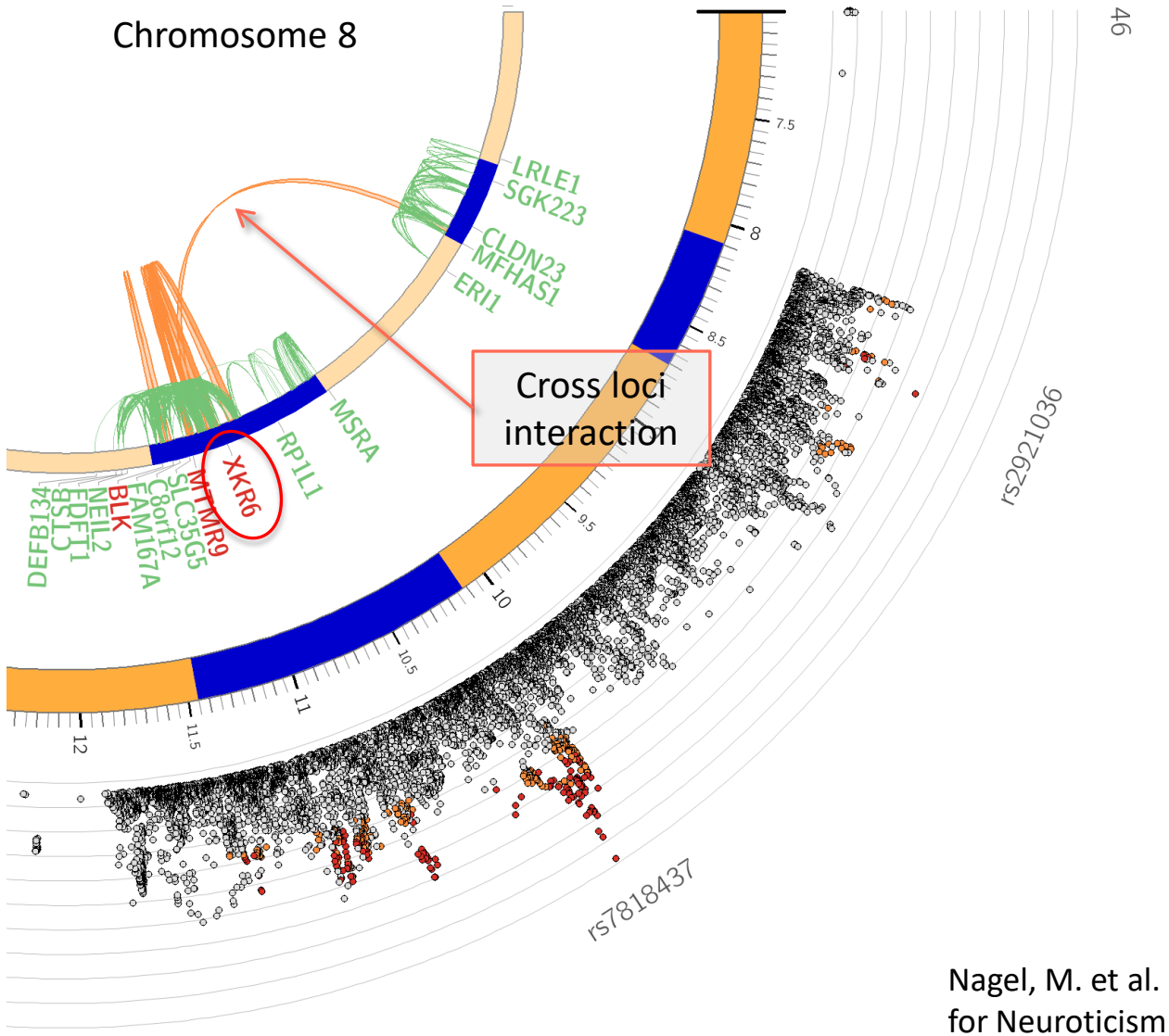
FUMA: circos plot



Chromatin interactions can map SNPs to distal genes.

Nagel, M. et al. Meta-Analysis of Genome-wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways. (2018) *Nat. Genet*.

FUMA: circos plot



Chromatin interactions can map SNPs to distal genes.

XKR6 is mapped by within and cross loci interactions and also mapped by eQTLs.

Nagel, M. et al. Meta-Analysis of Genome-wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways. (2018) *Nat. Genet*.

FUMA Practical

Shared file area:

Home/Danielle/FUMA/FUMApractical2023.docx