How do we go from genetic discoveries (from GWAS/WGS/WES) to mechanistic disease insight?

Part I – Functional annotation in risk loci

2021 Online International Statistical Genetics Workshop

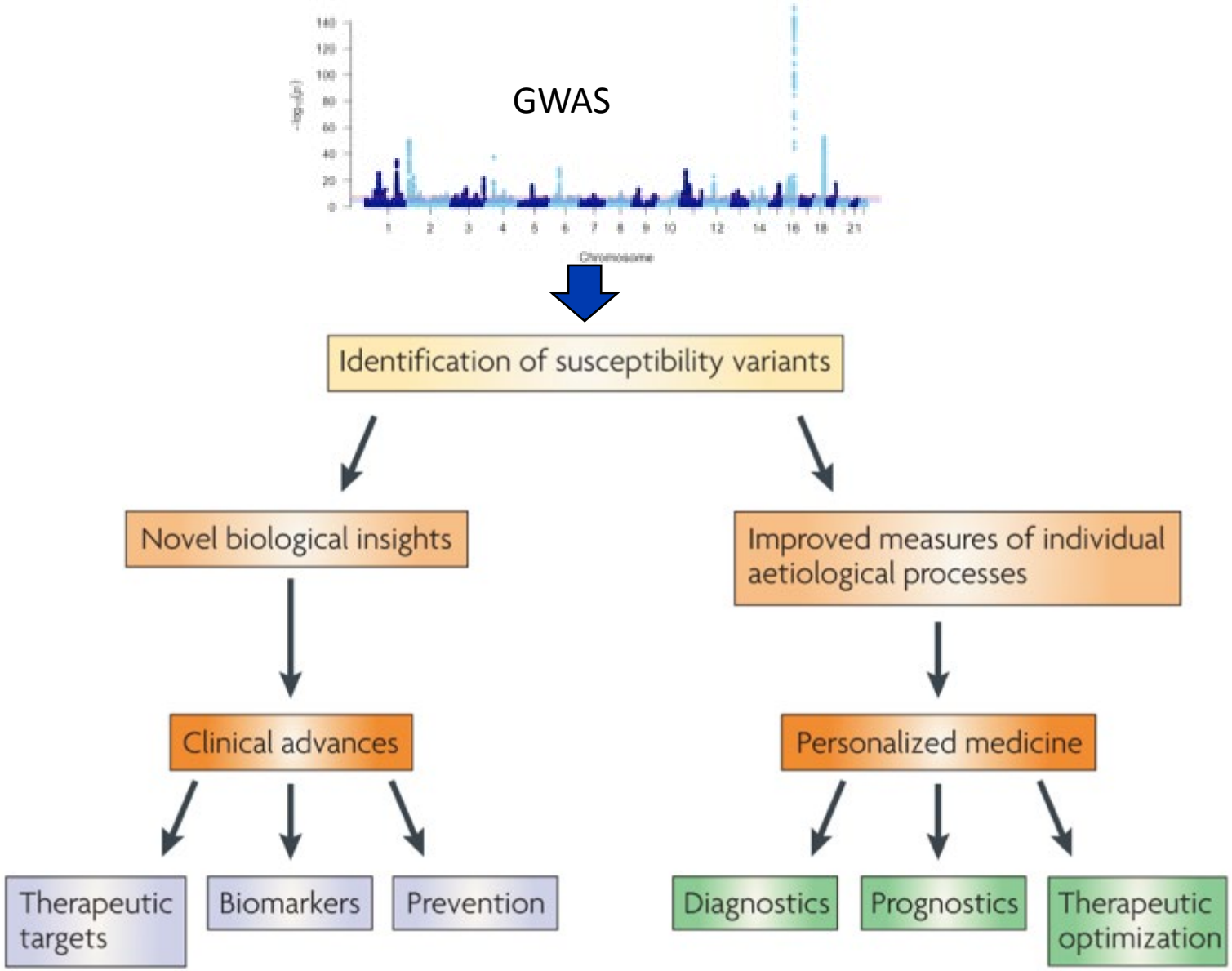
Danielle Posthuma | d.posthuma@vu.nl | @dposthu | https://ctg.cncr.nl

What have we learned so far?

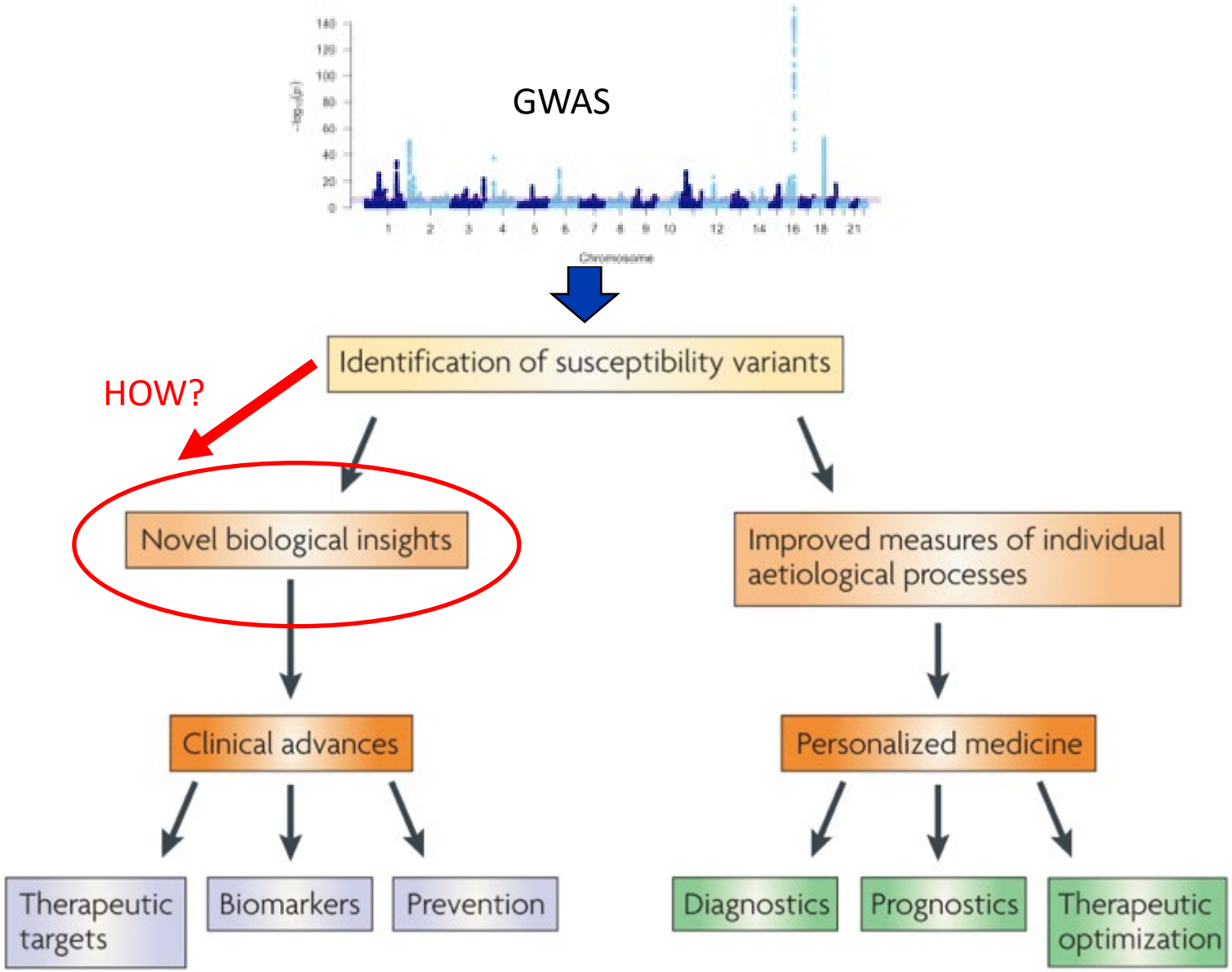
- ✓ Theory underlying genetic association
- ✓ Setting up a genome-wide association study
- ✓ Quality control for genetic datasets and analysis
- ✓ Conducting genetic association
- ✓ Several post-gwas analyses including SNP h^2, causal modeling, gSEM

Primary outcome of a genome-wide genetic association:

- Manhattan plot
- Summary statistics that include an effect estimate and significance of association per variant



McCarthy et al. *Nat. Rev. Gent.* (2008)



McCarthy et al. *Nat. Rev. Gent.* (2008)

How to gain mechanistic insight from genetic discoveries

Mendelian or monogenic disorders (influenced by one mutation in one gene)

- Segregation analysis (1970s – onwards) detected several genes co-segregating with disease
- For each disease a mutation in one gene is sufficient to express that disease
- Functional experimentation on these genes involved e.g. knock-out models to investigate that gene's function
- This has been successful for e.g. PKU, Huntington's disease, breast cancer.
- Any mechanistic insight guides treatment development

How to gain mechanistic insight from genetic discoveries

Polygenic disorders (influenced by 100's of variants each of small effect)

- GWAS (2006s – onwards) detected several genetic loci associated with diseases that are polygenic
- For each disease a single genetic variant is not sufficient to express that disease, instead 100's of variants cumulatively increase risk for disease
- Detected loci contain 100's of variants, sometimes no genes are implicated
- Functional experimentation on these variants is not straightforward, mechanistic insight is not easily obtained for polygenic traits

GWAS hits for polygenic traits often not directly useful for functional follow-up

4 issues:

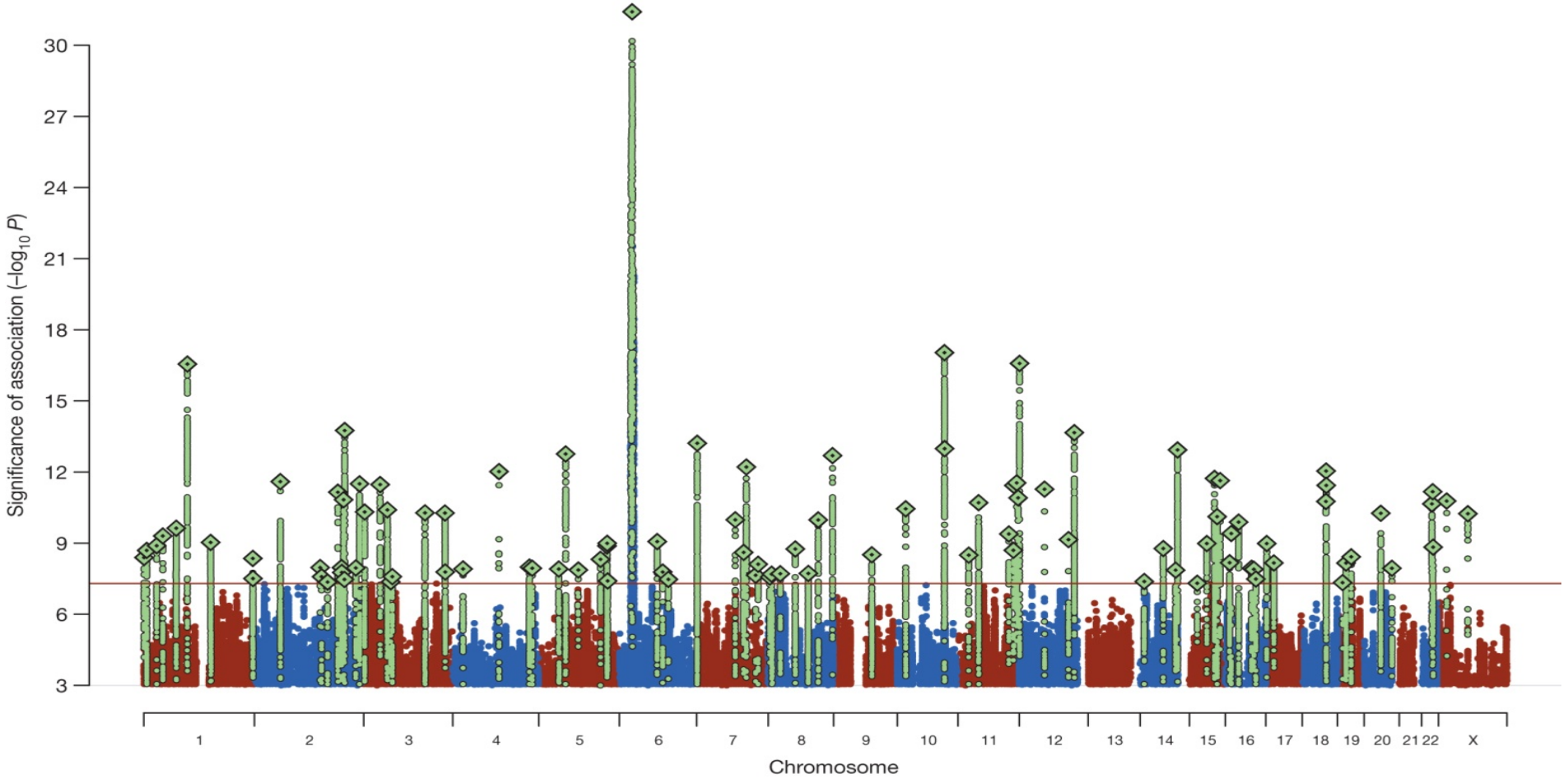
1. GWAS hits for polygenic traits mostly outside genes, or in non-coding genic regions, with likely regulatory functions that are currently unknown
2. GWAS hits for polygenic traits have small effects, making them unsuitable for small-scaled/under-powered functional studies
3. SNPs are correlated (LD) which complicates pinpointing ‘the’ causal SNP
4. There are 100’s of genes involved in polygenic traits – a single gene will not provide the whole picture

GWAS hits for polygenic traits often not directly useful for functional follow-up

4 issues:

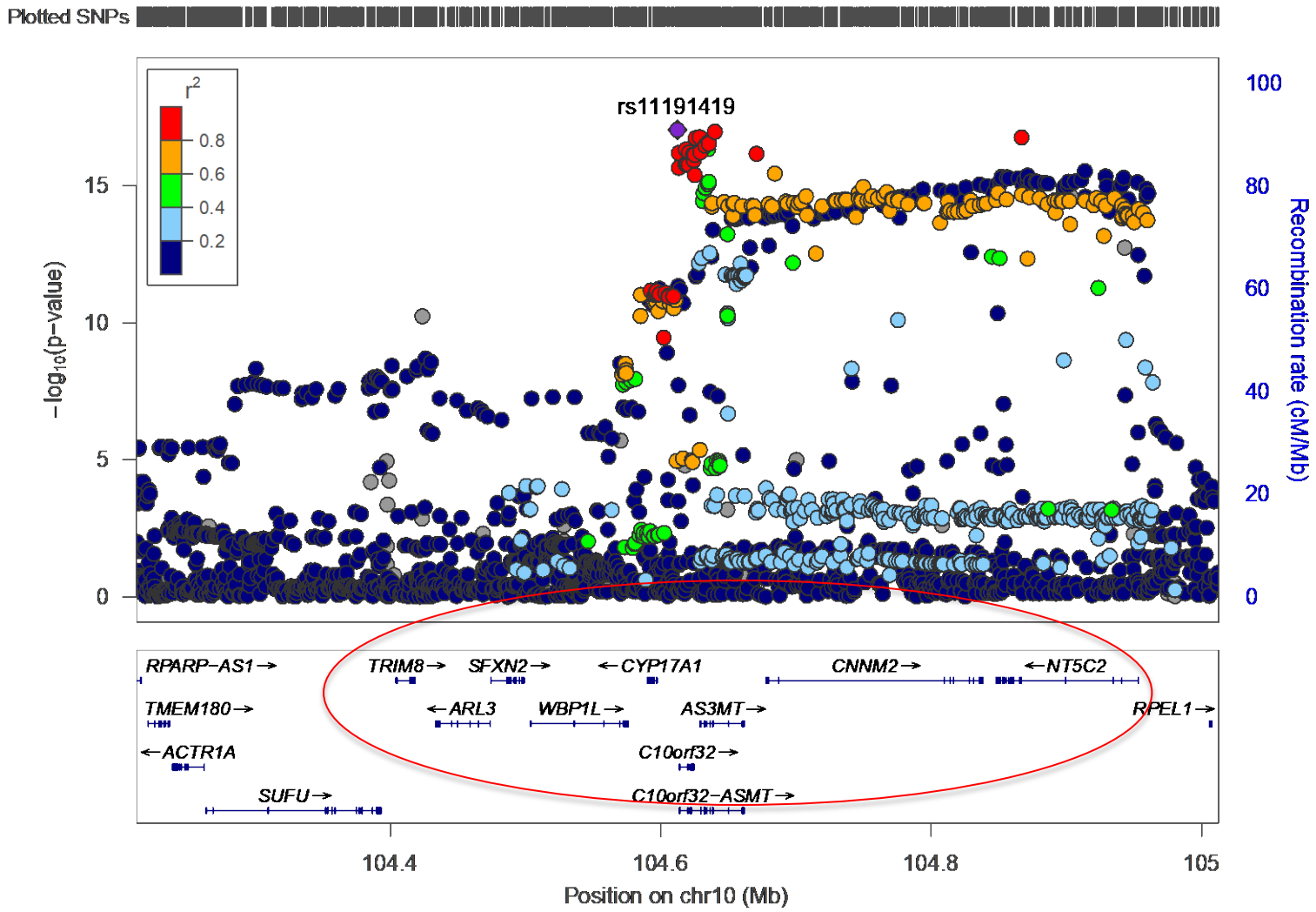
1. GWAS hits for polygenic traits mostly outside genes, or in non-coding genic regions, with likely regulatory functions that are currently unknown
2. GWAS hits for polygenic traits have small effects, making them unsuitable for small-scaled/under-powered functional studies
3. SNPs are correlated (LD) which complicates pinpointing ‘the’ causal SNP
4. There are 100’s of genes involved in polygenic traits – a single gene will not provide the whole picture

GWAS result



S Ripke et al. Nature (2014)

Zooming in on a locus



SNPs are correlated (LD) which complicates pinpointing ‘the’ causal SNP

The genotypes on SNPs close to each other tend to be correlated due to linked segregation

Therefore, statistical associations will be picked up with all SNPs that are correlated with the causal SNP

SNPs are correlated (LD) which complicates pinpointing ‘the’ causal SNP

How to prioritize most likely causal SNPs/genes?

- Take the gene closest to the most significant SNP

Often, but not always seems to be a good guess

- Statistical fine-mapping

Model the known correlation structure against the observed pattern of association values to pinpoint the most likely causal SNPs assuming N causal SNPs, can be integrated with functional information (tools FINEMAP, PAINTOR)

- Functional annotation

Variants with a known effect on transcription or protein structure are more likely to be causal than non-functional ones (tools FUMA, VEP, ANNOVAR)

Functional categories of SNPs

For a SNP to be potentially causal, it needs to affect the gene, either via structure or via regulatory functions

-> Step 1 after GWAS: annotate associated SNPs with known functions

- **Protein Coding**
  - SNPs in exonic regions may alter protein structure and/or function e.g nonsense SNPs or missense SNPs
- **Splicing Regulation**
  - SNPs in splice sites may disrupt splicing regulation, resulting in exon skipping or intron retention
  - They can also interfere with alternative splicing regulation by changing exonic splicing enhancers or silencers.
- **Transcriptional Regulation**
  - SNPs in transcription regulatory regions (e.g. transcription factor binding sites, CpG islands, microRNAs, etc.) can alter binding sites, and thus disrupt proper gene regulation.
- **Post-Translational Modification**
  - SNPs in protein-coding regions may alter post-translational modification sites, interfering with proper posttranslational modification.

Interpreting GWAS risk loci

- Are there functional variants in the GWAS risk loci?

  E.g. nonsynonymous coding SNPs

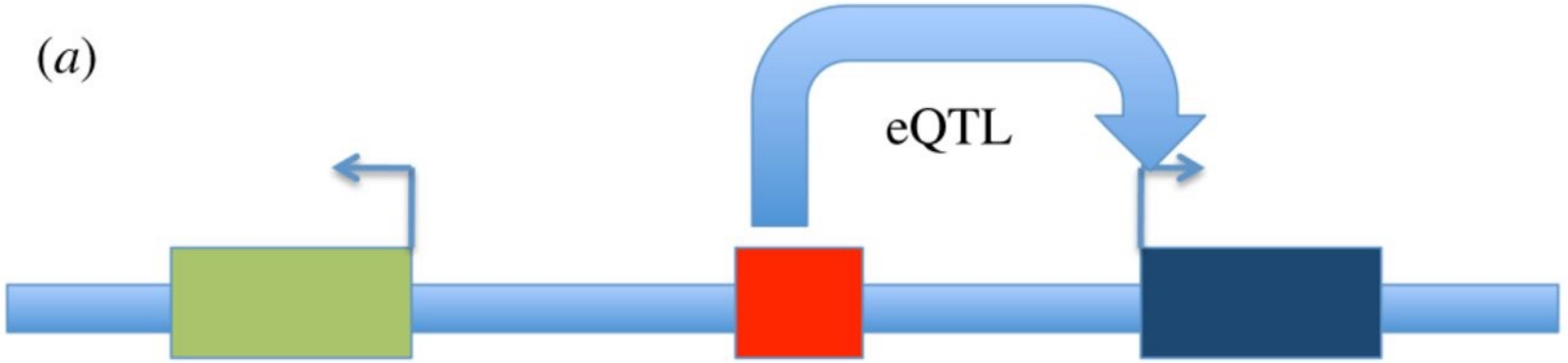
- Are there SNPS that are likely to be deleterious?

  E.g. SNPs with high (>~10) ‘CADD’ scores

- Are there SNPs likely to have regulatory effects on genes?

  E.g. SNPs with low RegulomeDB scores, or eQTLs (SNPs previously associated with differences in RNA levels), SNPs that are know to physically overlap with promoter regions when the DNA is folded, via HiC interaction

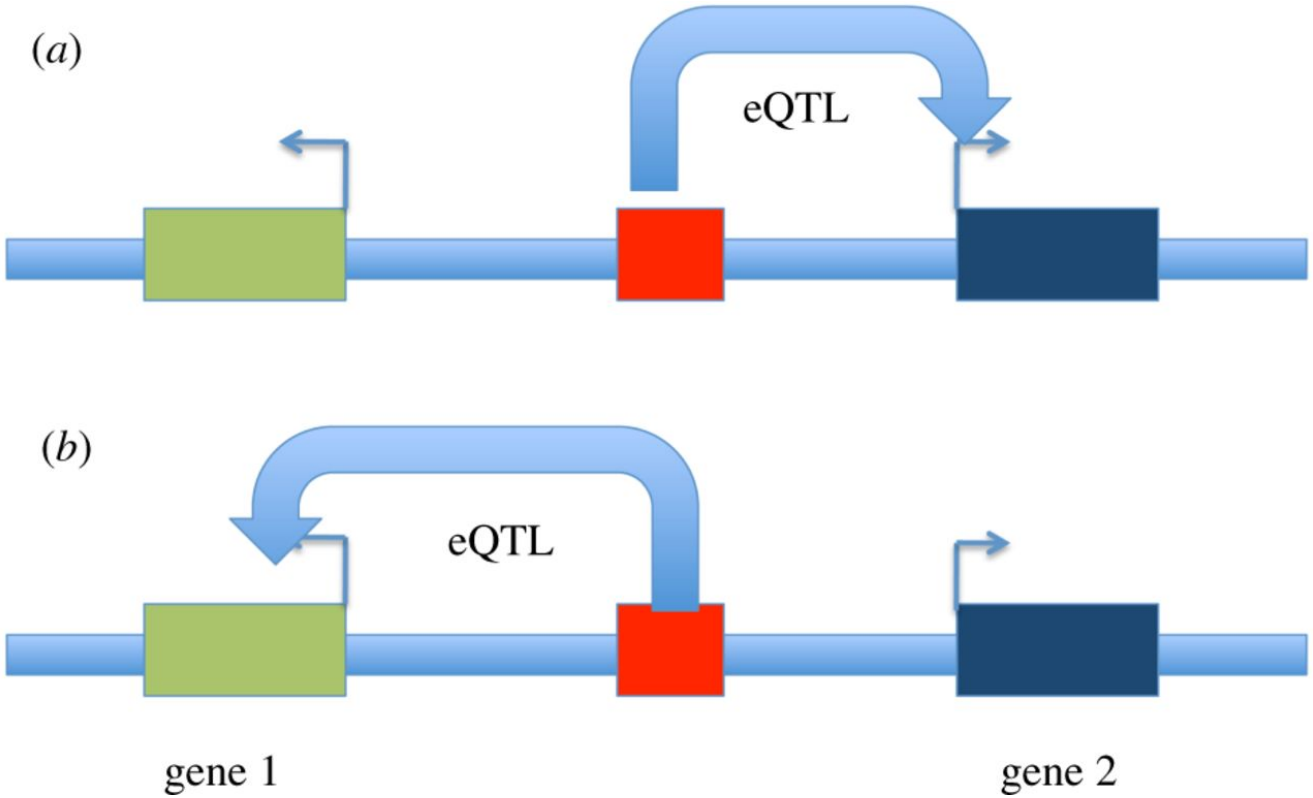
Expression QTLs



Alexandra C. Nica, and Emmanouil T. Dermitzakis Phil. Trans. R. Soc. B 2013;368:20120362

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

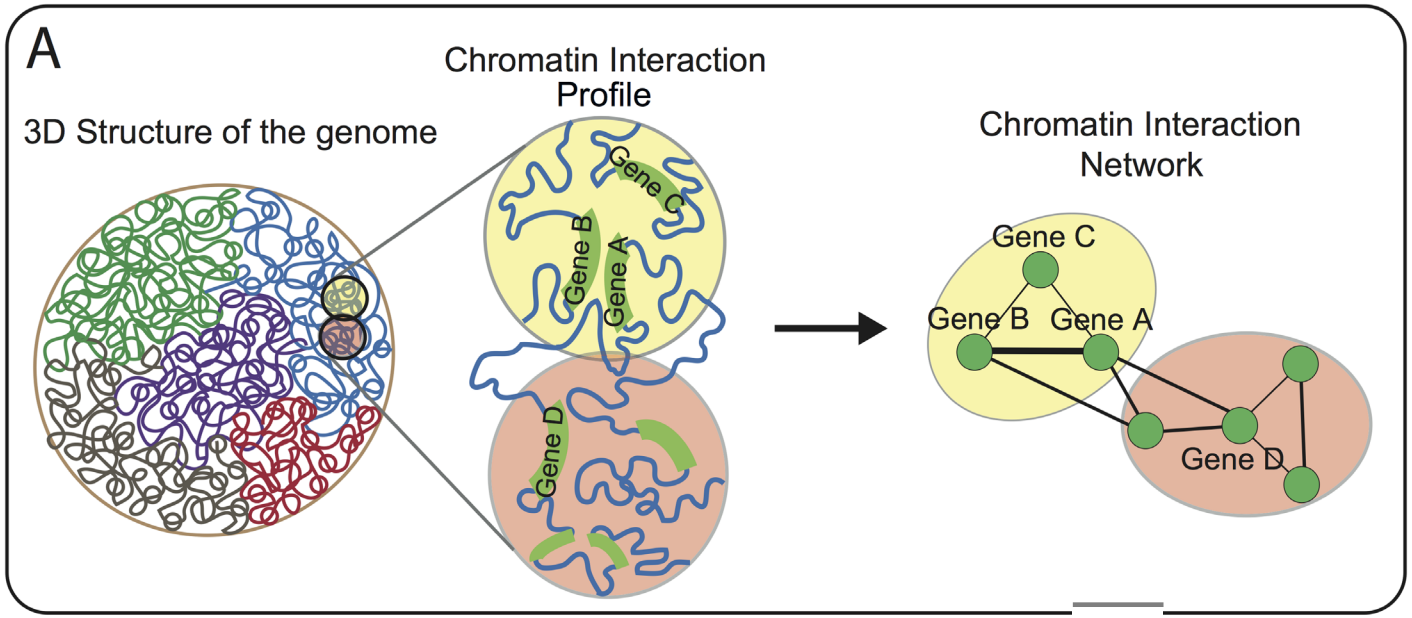
The same regulatory regions and variant could be an eQTL for gene 2 in (a) tissue 1 and for gene 1 in (b) tissue 2, suggesting that limited interrogation of tissues would be misleading for the biological signal underlying disease.



Alexandra C. Nica, and Emmanouil T. Dermitzakis Phil. Trans. R. Soc. B 2013;368:20120362

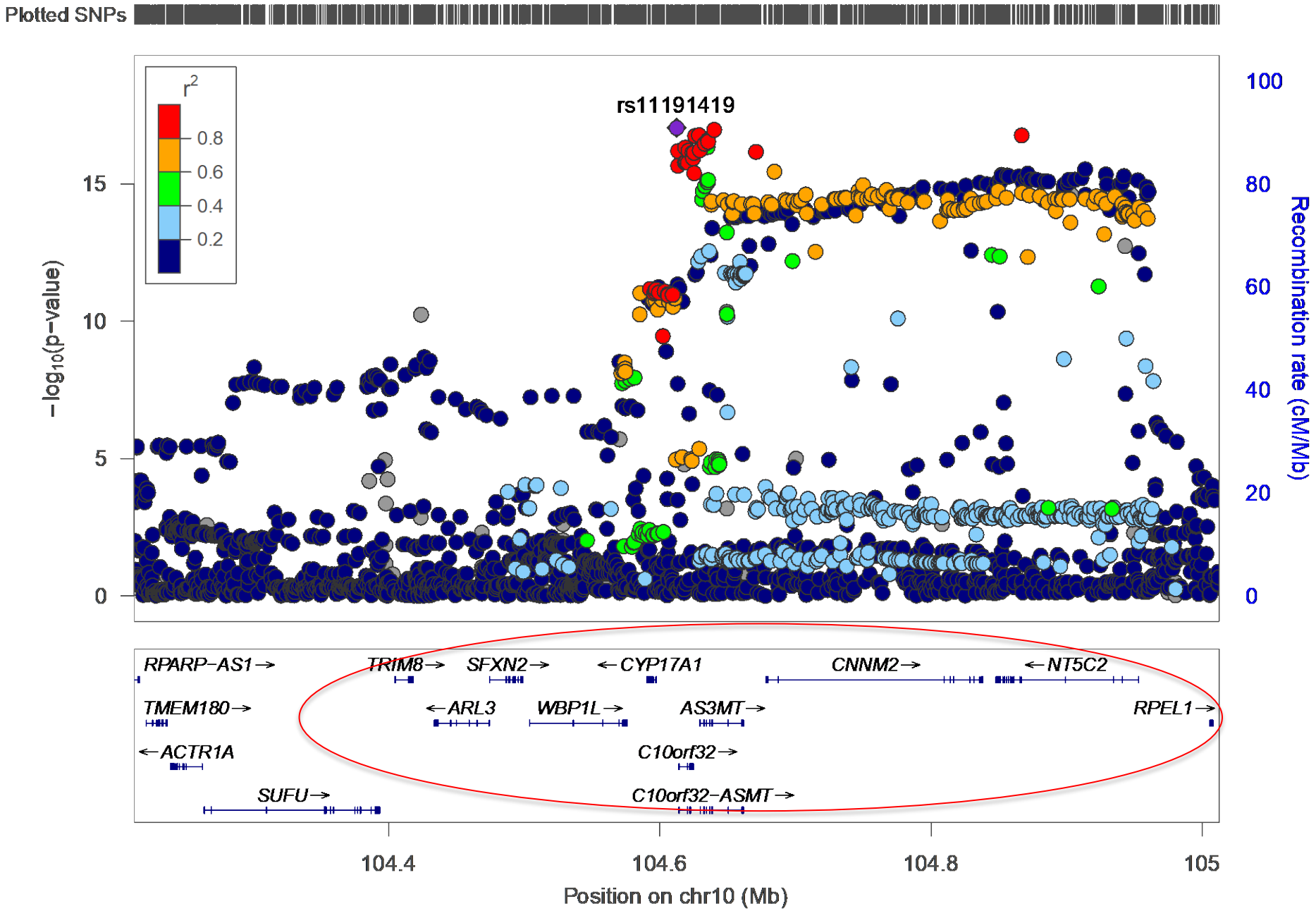
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Chromatin (HiC) interaction

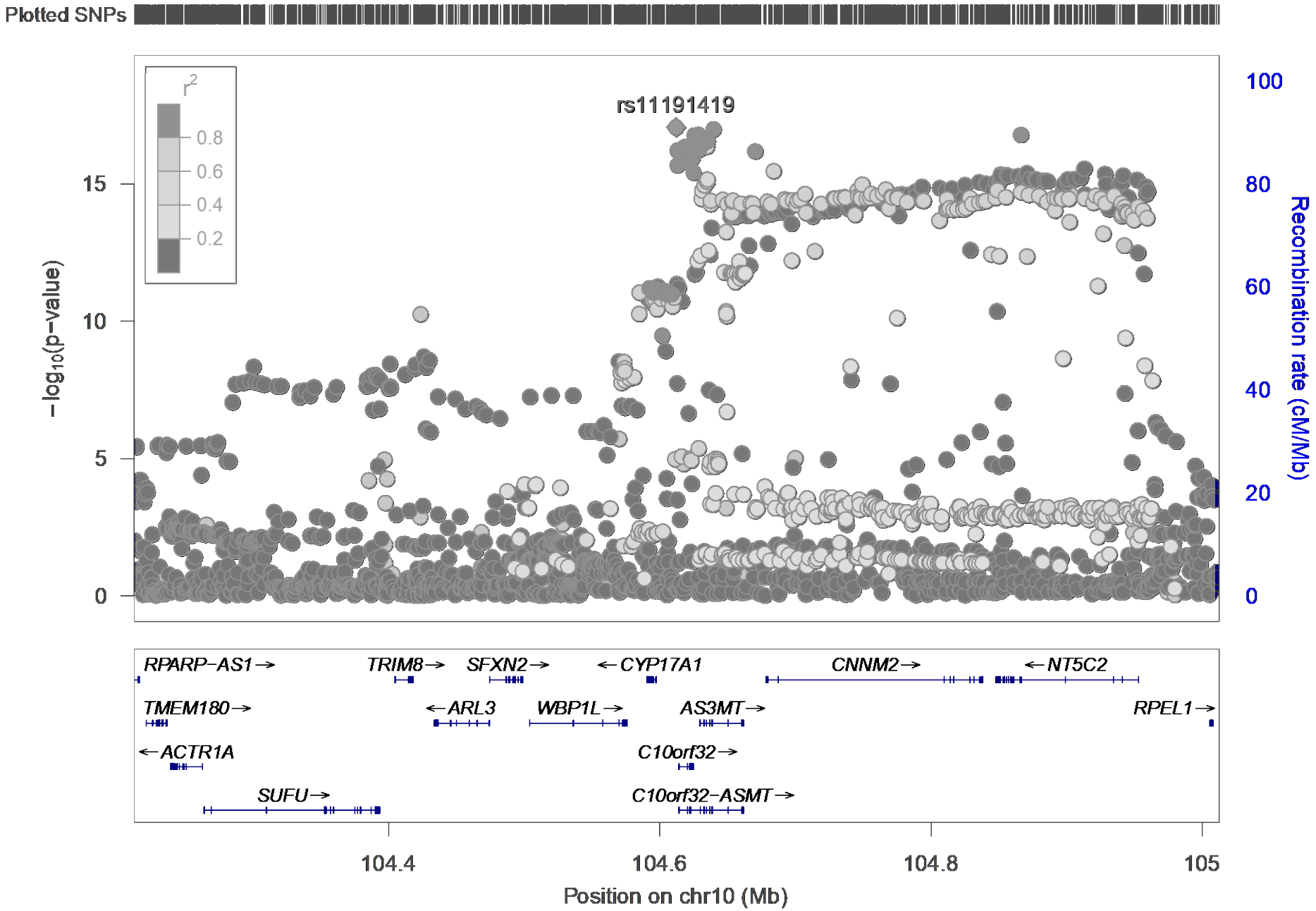


Babaei et al Plos Comp Biol, 2015

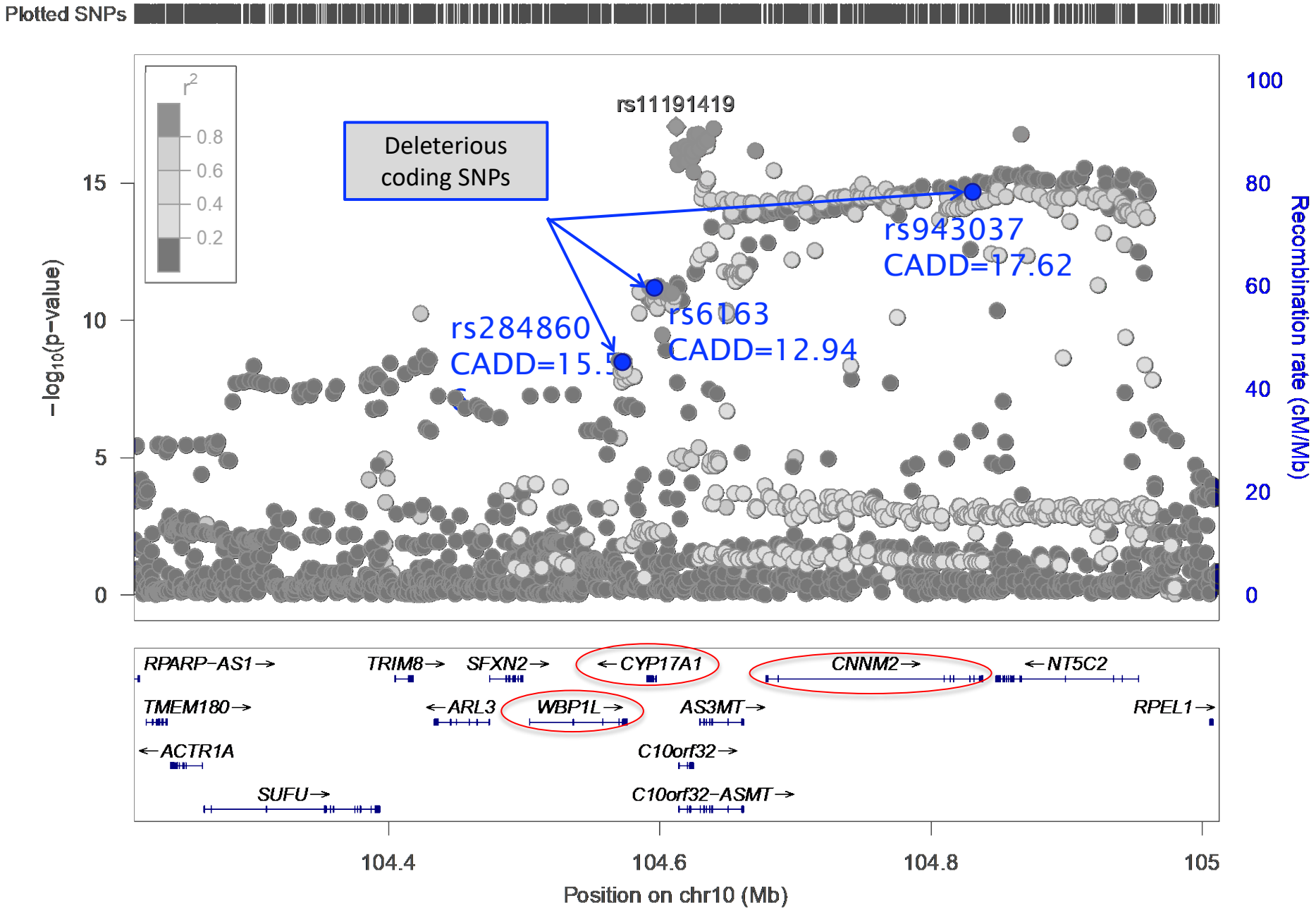
GWAS risk locus



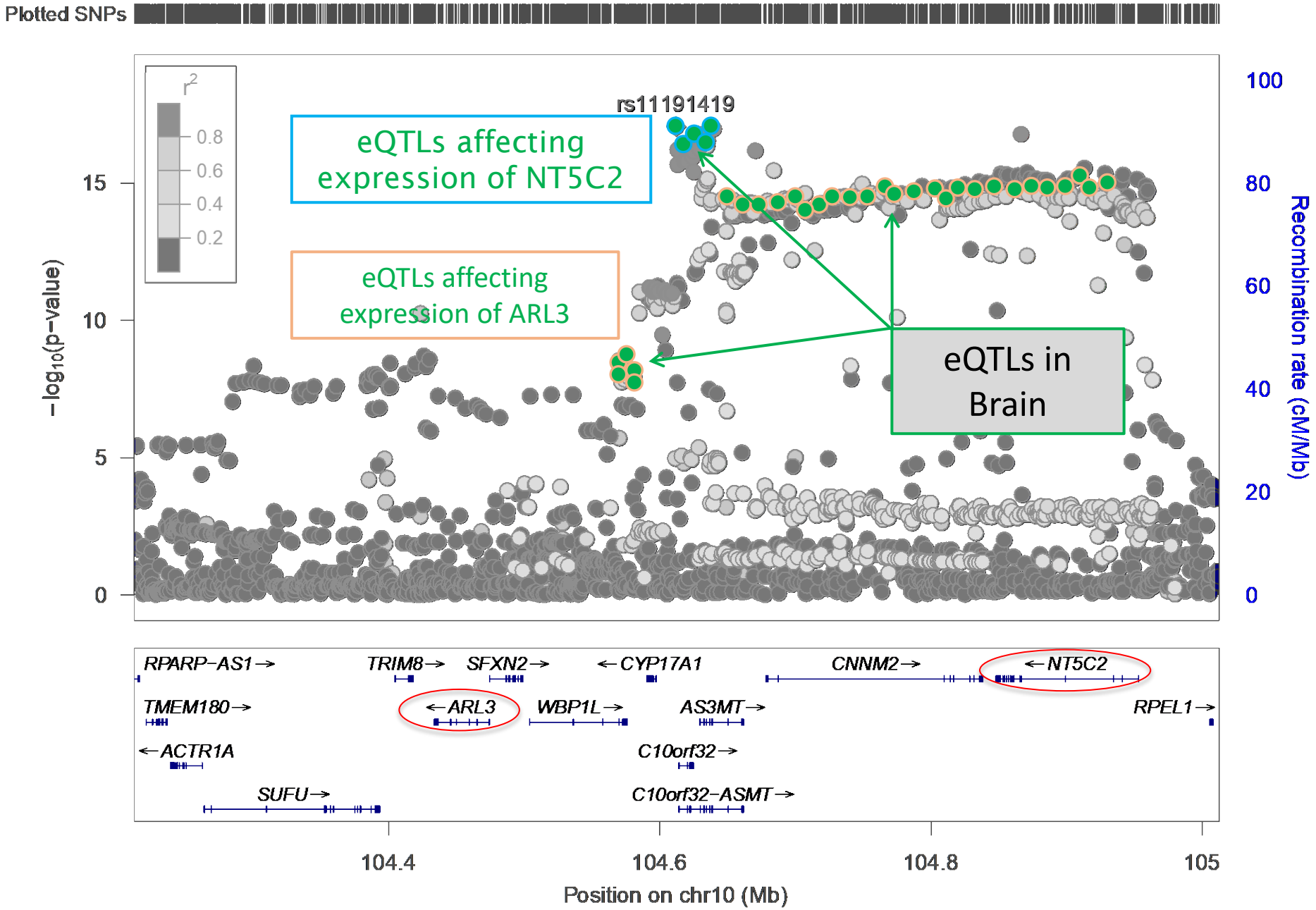
GWAS risk locus



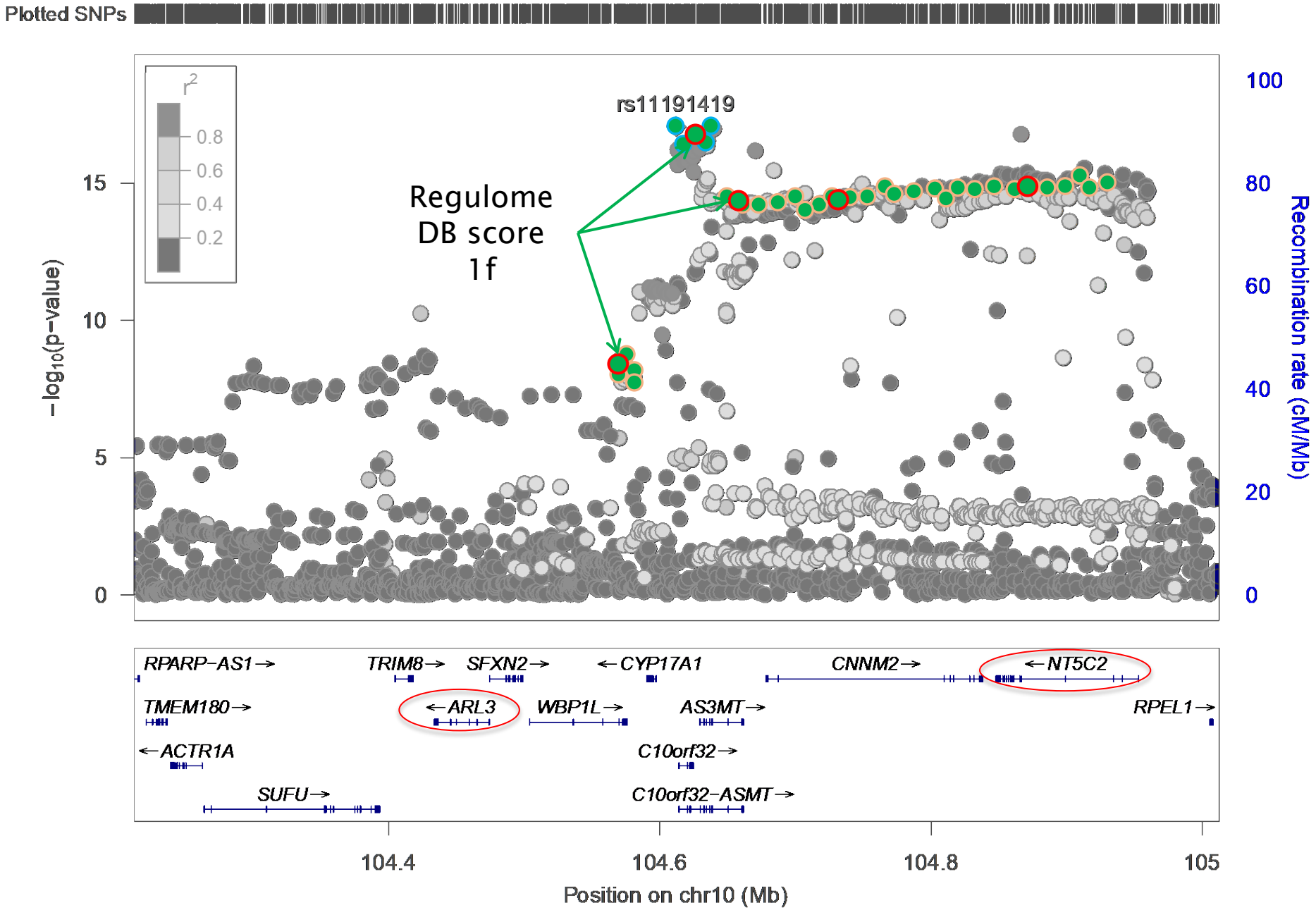
GWAS risk locus



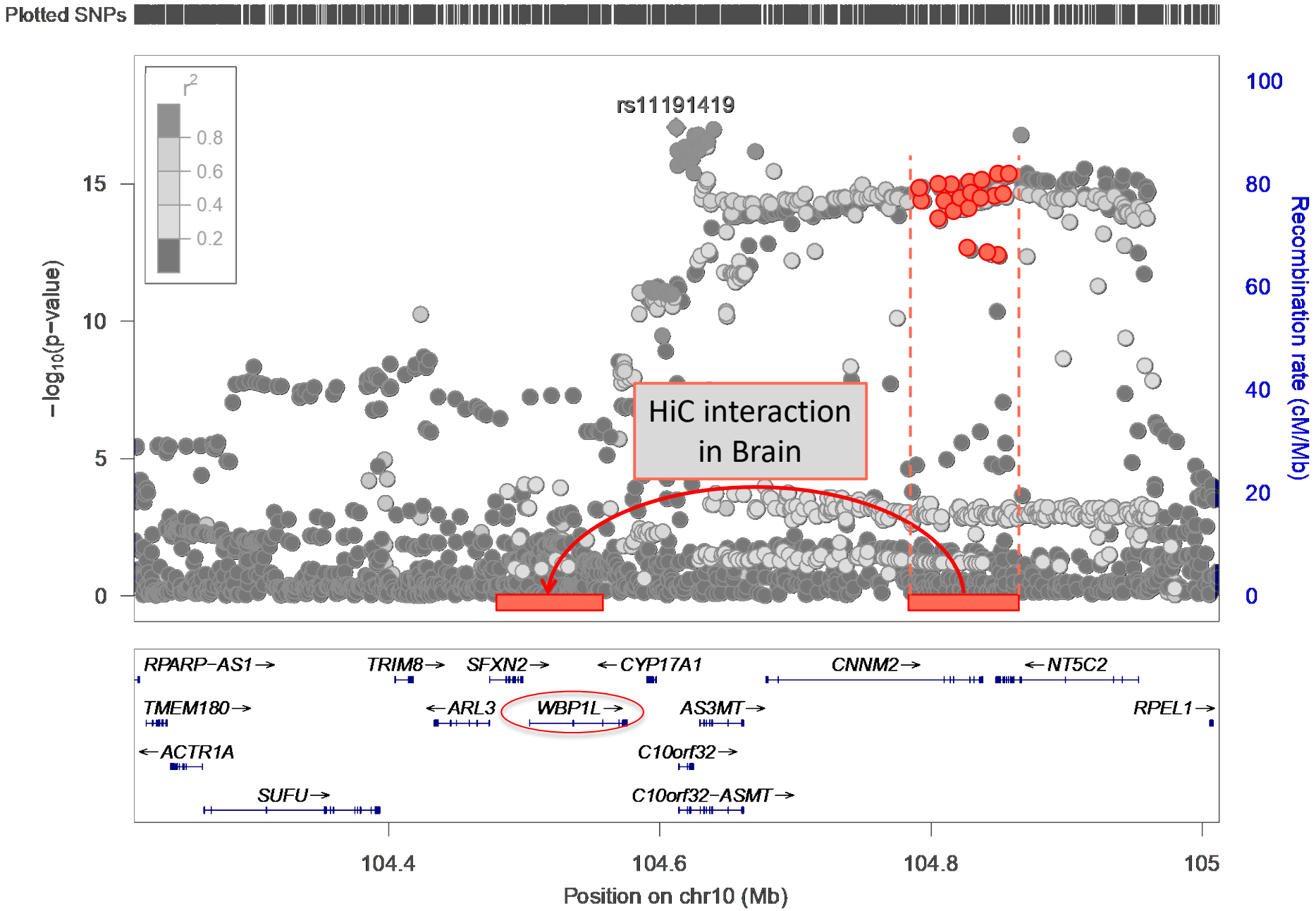
GWAS risk locus



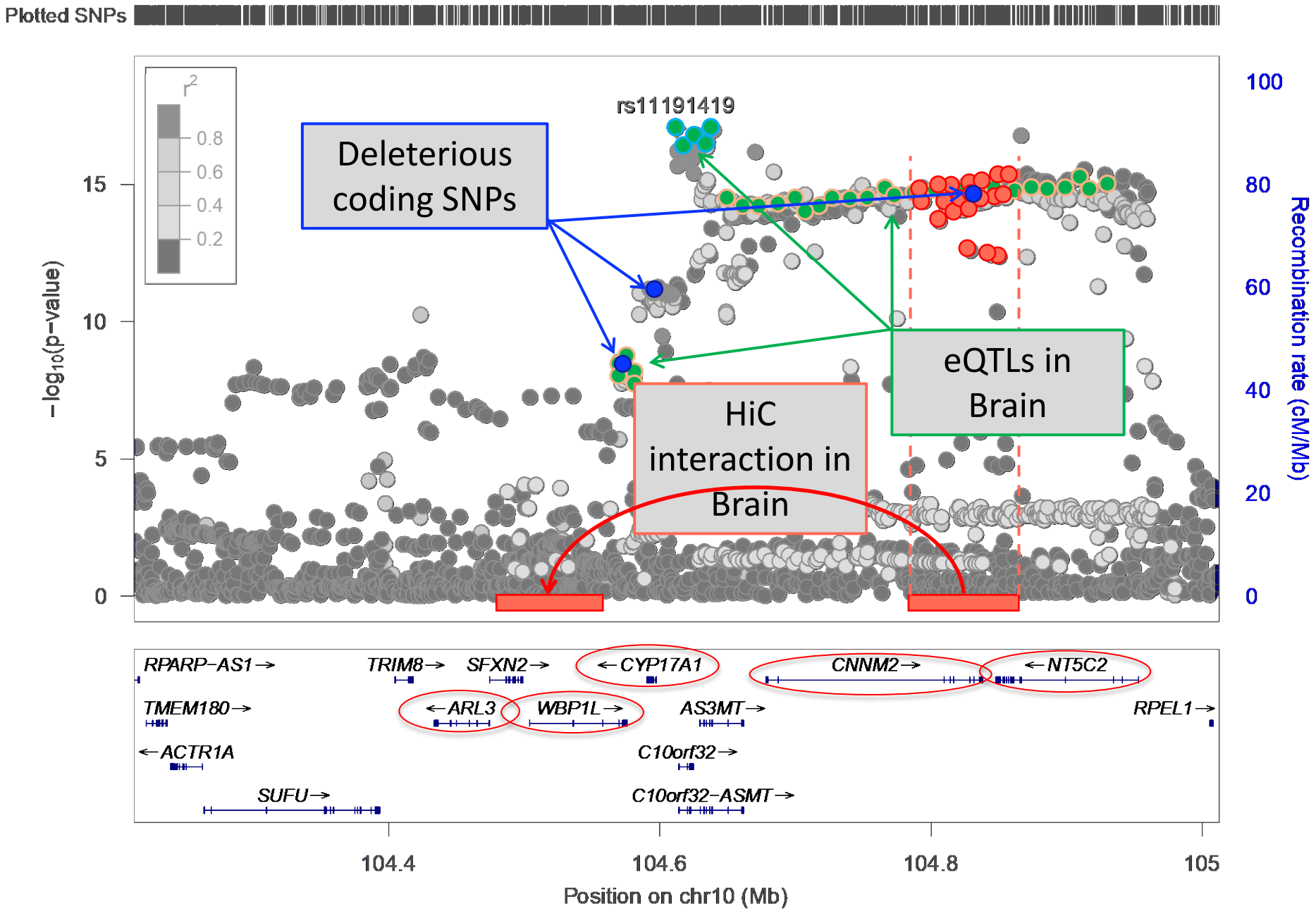
GWAS risk locus



GWAS risk locus



GWAS risk locus



Interpreting GWAS results

1. Linkage disequilibrium (LD)
   Identify all SNPs which are in LD of significant hits.

PLINK

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

ANNOVAR SnpEff VeP

3. Functional annotation
   Deleteriousness, regulatory elements and epigenetic da

PsychENCODE Consortium

CADD

GTEx Portal

BBMRI.nl
Biobanking and BioMolecular resources Research Infrastructure The Netherlands

ENCODE

RegulomeDB

HaploReg

ROADMAP epigenomics PROJECT

4. Functional analyses of genes
   Tissue specific expression, gene set analyses

MSigDB
Molecular Signatures Database

WikiPathways
Pathways for the People

KEGG
Kyoto Encyclopedia of Genes and Genomes

GEO
Gene Expression Omnibus

Interpreting GWAS results

1. Linkage disequilibrium (LD)
   Identify all SNPs which are in LD of significant hits.

PLINK

2. Var
   Fu

AN

3. Fu
   D

CA

4. Ful
   Tissue specific expression; gene set analyses

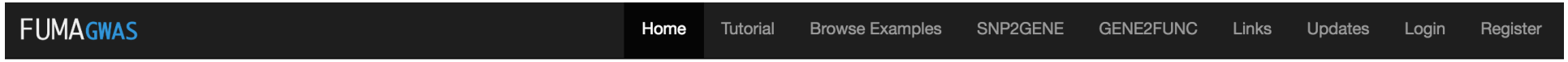
Multiple databases
Multiple software
Multiple steps
Reformatting of data

Time-consuming + error prone

MSigDB
Molecular Signatures Database

FUMA: Functional Mapping and Annotation of genetic associations

Available at http://fuma.ctglab.nl



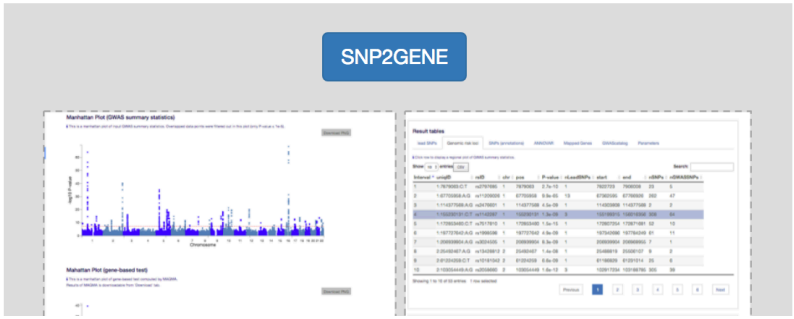
FUMA GWAS

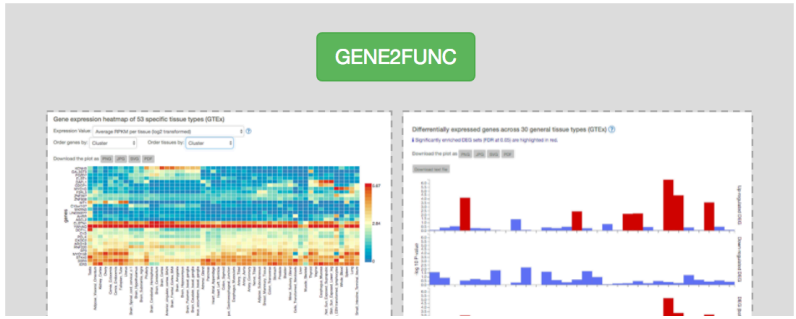
Functional Mapping and Annotation of Genome-Wide Association Studies

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).

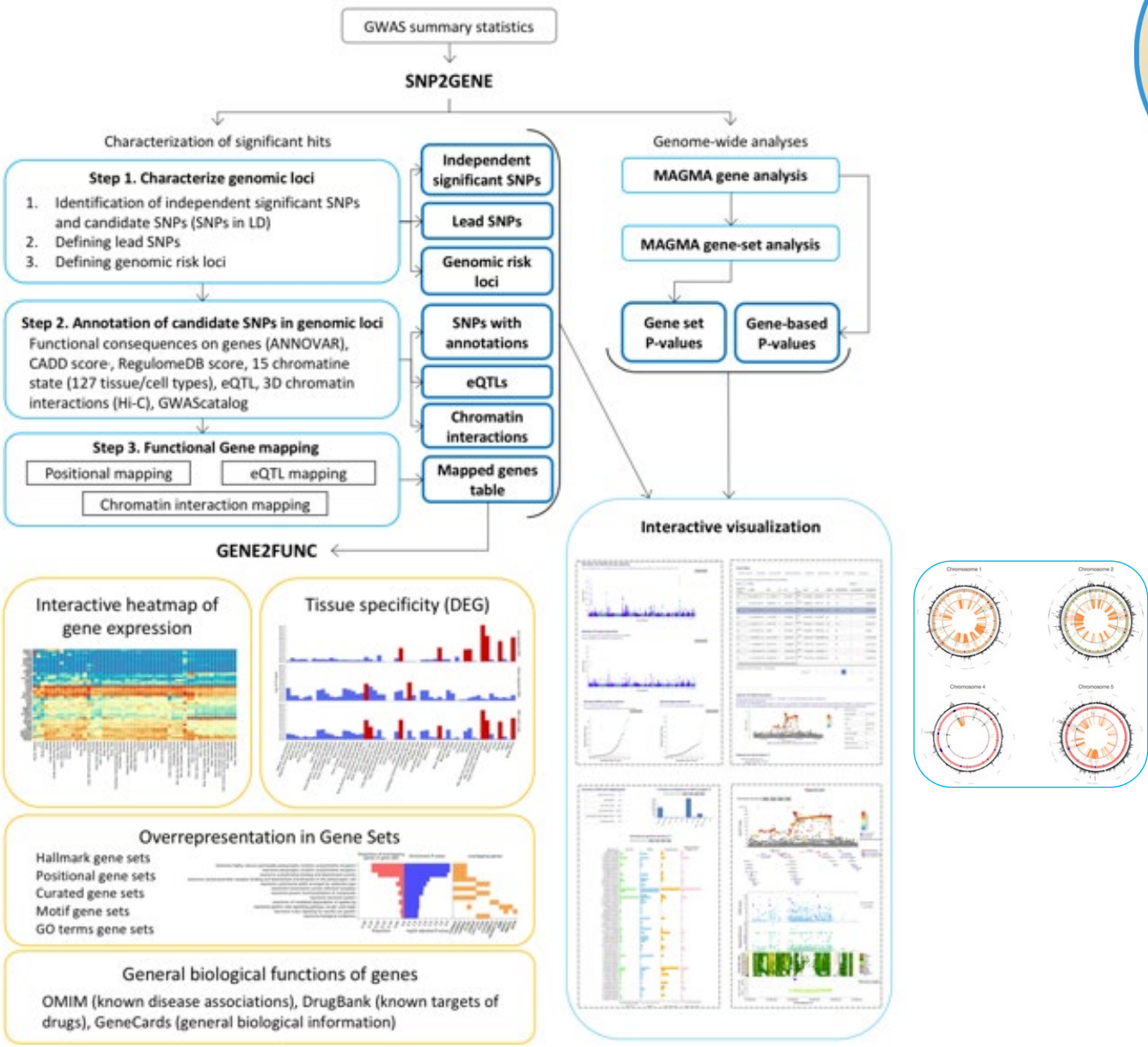




FUMA developed by Kyoko Watanabe

fuma.ctglab.nl

Watanabe K, Taskesen, van Bochoven
Posthuma D. 2017 NatComm



So far…

Locus based interpretation -> prioritization of SNPs and genes within a locus

But -100’s of loci -> we also need to interpret across loci

GWAS hits for polygenic traits often not directly useful for functional follow-up

4 issues:

1. GWAS hits for polygenic traits mostly outside genes, or in non-coding genic regions, with likely regulatory functions that are currently unknown
2. GWAS hits for polygenic traits have small effects, making them unsuitable for small-scaled/under-powered functional studies
3. SNPs are correlated (LD) which complicates pinpointing ‘the’ causal SNP
4. There are 100’s of genes involved in polygenic traits – a single gene will not provide the whole picture

SNP annotation implicates genes – after this: look for convergence

- Explore gene functions
- Explore pathway enrichment of implicated genes
- Explore in which tissue genes are expressed
- Explore which cell types are indicated

Interpreting GWAS outcomes

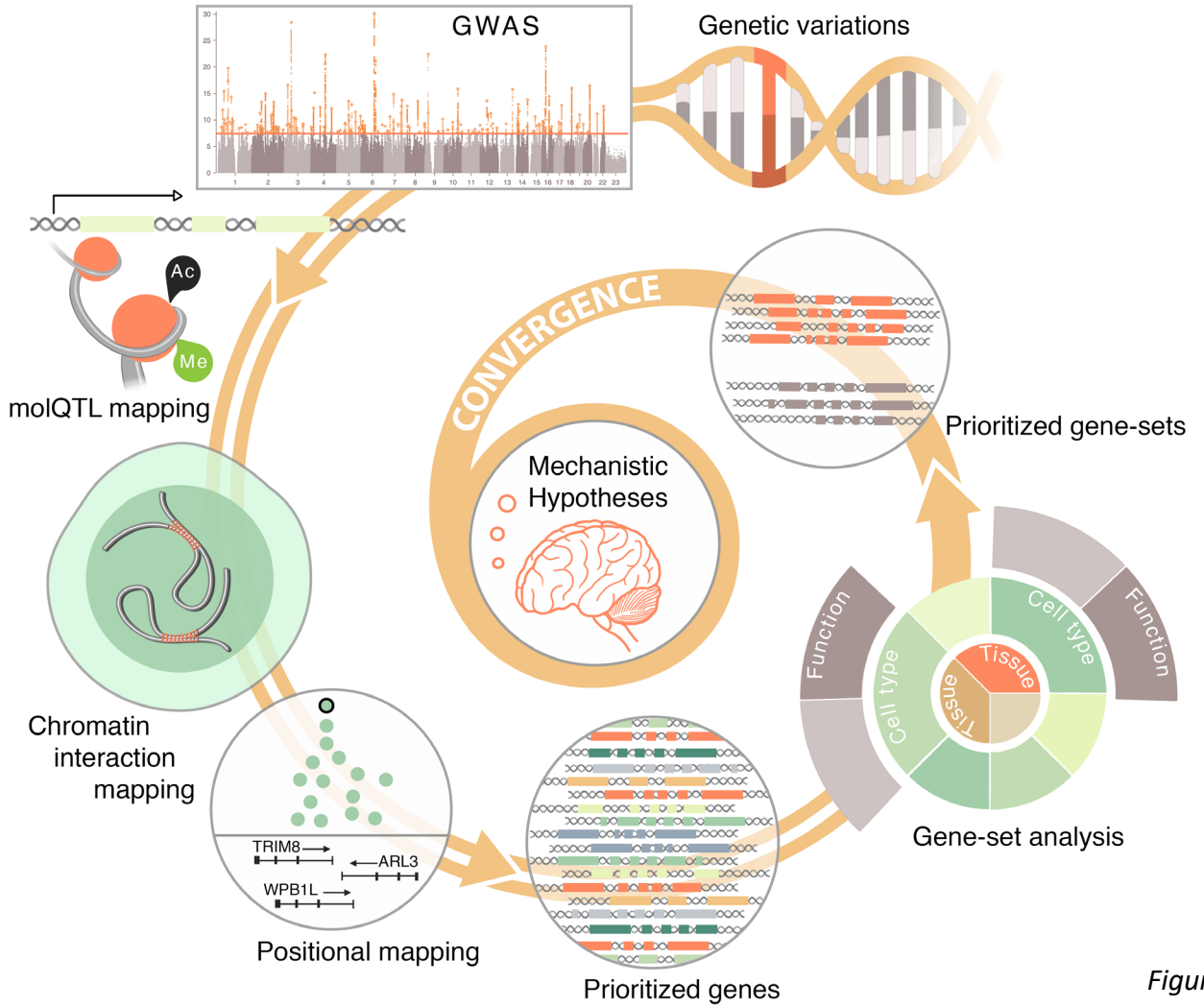


Figure from Uffelmann & Posthuma, Biol Psychiatry, 2020