Post GWAS analyses: Characterising GWAS loci

Teresa Ferreira and Nilufer Rahmioglu

Wellcome Centre for Human Genetics/Big Data Institute

Acknowledgment: Slides are courtesy of Prof. Andrew Morris

Learning Objectives

- How to characterize GWAS loci to identify causal variants and provide insights into biology.
  - Identify multiple causal variants at one locus
  - Localisation of causal variants: Credible sets
  - Trans-ethnic fine-mapping
  - fGWAS
  - eQTL Colocalisation
  - LD Score Regression

Characterising GWAS loci

- GWAS have been successful in identifying loci contributing to complex disease.
- Loci typically characterised by common variants that are in strong LD with each other, and consequently have similar strength of association.
- How can we identify the causal variant(s)?
- Provides insight into upstream biology (e.g. causal variants tend to map to enhancers in a specific tissue) and downstream biology (e.g. effect of variant on disease is mediated through a specific gene).

Do we have evidence of multiple causal variants at a locus?

- Typical to first dissect association signals at GWAS loci that reflect different underlying causal variants.
- Can be achieved through conditional analysis: include genotypes at SNP with strongest association signal as a covariate.
- Iterative “forward” selection: add SNP with strongest association signal as additional covariate until there is no residual association.
- Approximate conditional analyses implemented in GCTA software:
  - Individual level genotype data not required: association summary statistics.
  - Reference genotype dataset for LD: correlation between test statistics in joint model.
  - Implements backward selection to identify index SNPs for each distinct association signal.

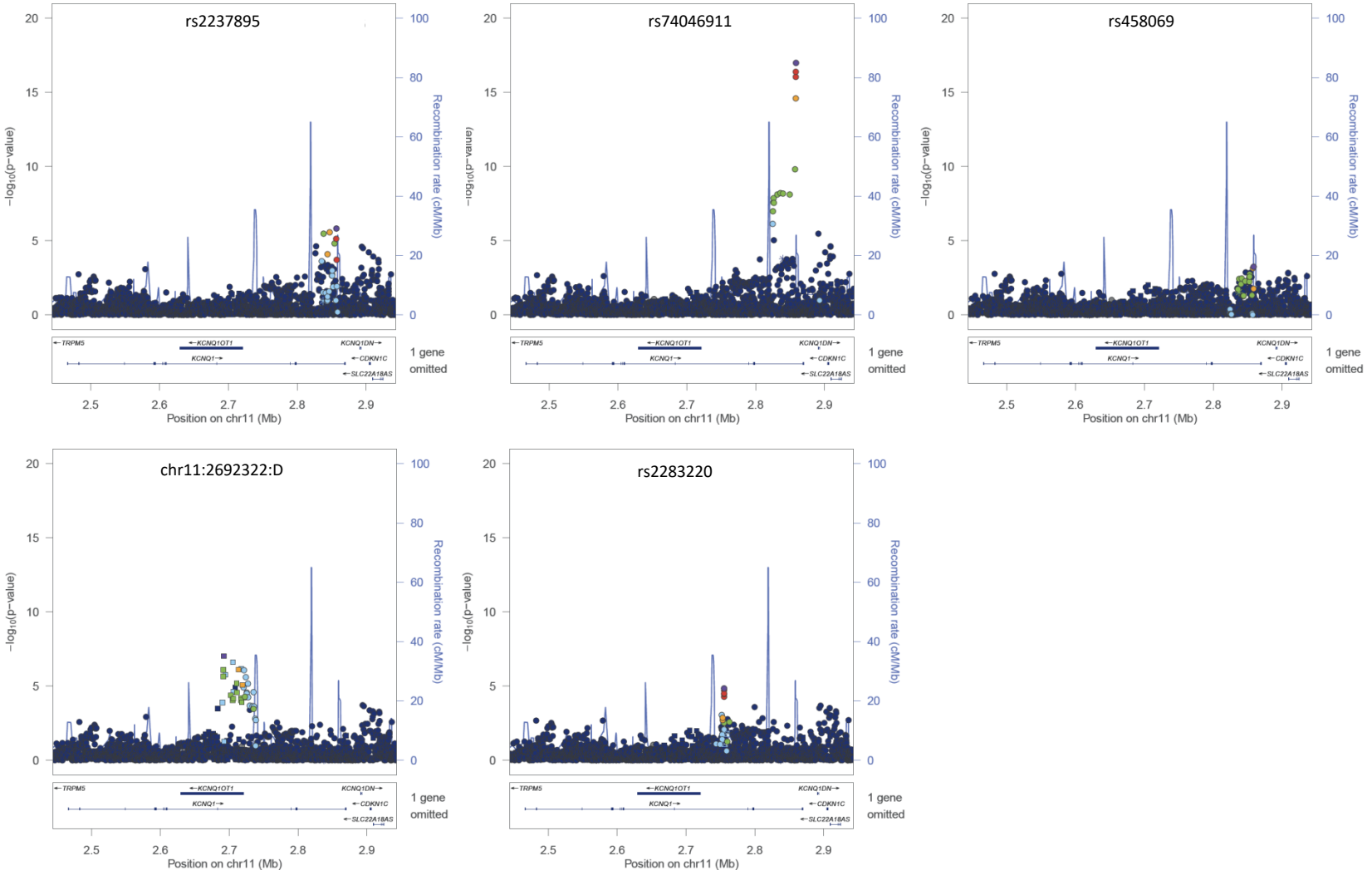
Dissecting T2D association signals

- Total of 27,206 T2D cases and 57,574 controls from 23 studies of European ancestry, genotyped with the Metabochip.
  - Custom iSELECT array containing ~195K SNPs, designed to support large-scale follow-up of putative associations for T2D and other metabolic and cardiovascular traits.
  - High-density coverage of variation from 1000 Genomes Project pilot data in 180 fine-mapping regions overlapping 39 established loci for T2D susceptibility.
- Evaluate the evidence for multiple signals of association at established T2D susceptibility loci.

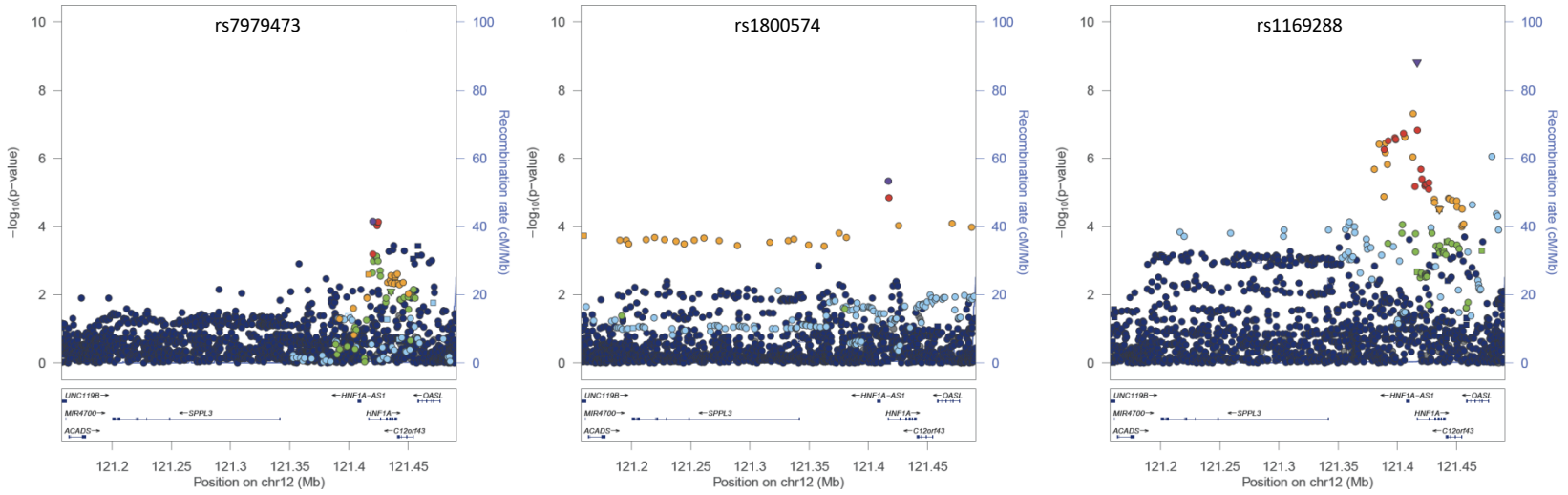
Dissecting T2D association signals

- Approximate conditional analysis undertaken using the GCTA software using 3,298 cases and 3,708 controls from GoDARTS as reference.
- *In silico* replication of association signals in an additional 19,662 T2D cases and 115,140 controls of European ancestry (combined meta-analysis $p<10^{-5}$).
- Confirmation of association signals through exact conditional analysis.

Five signals of association at *KCNQ1*



Multiple signals of association at six additional loci



- Conditional analyses revealed three signals of association mapping to *HNF1A*.
- Two signals of association each identified at *CDKN2A-B*, *DGKB*, *MC4R*, *GIPR* and *HNF4A*.

Localisation of causal variants

- Evaluate fine-mapping resolution on basis of statistical evidence of association by construction of “credible sets” of variants in each signal.
- Posterior probability of “causality” for each variant.
- Identify smallest set of variants that account for 99% of the probability of causality: 99% credible set.
- Smaller credible sets (number of variants and/or genomic interval covered) correspond to greater fine-mapping resolution.

Posterior probability of causality

- For each association signal, we require a Bayes' factor in favour of association for each variant.
  - Can be obtained directly from SNPTEST.
  - Can be approximated on the basis of association summary statistics via Wakefield approach:

$$\Lambda_j = \sqrt{\frac{V_j}{V_j+\omega}} \exp\left[\frac{\omega \beta_j^2}{2V_j(V_j+\omega)}\right],$$

  - Effect size β and corresponding variance V.
  - Prior variance of effect size ω: typically taken to be 0.04 for binary traits.
- Posterior probability of causality then given by:

$$\pi_{Cj} = \frac{\Lambda_j}{\Sigma_k \Lambda_k},$$

- Can incorporate prior probability of causality.

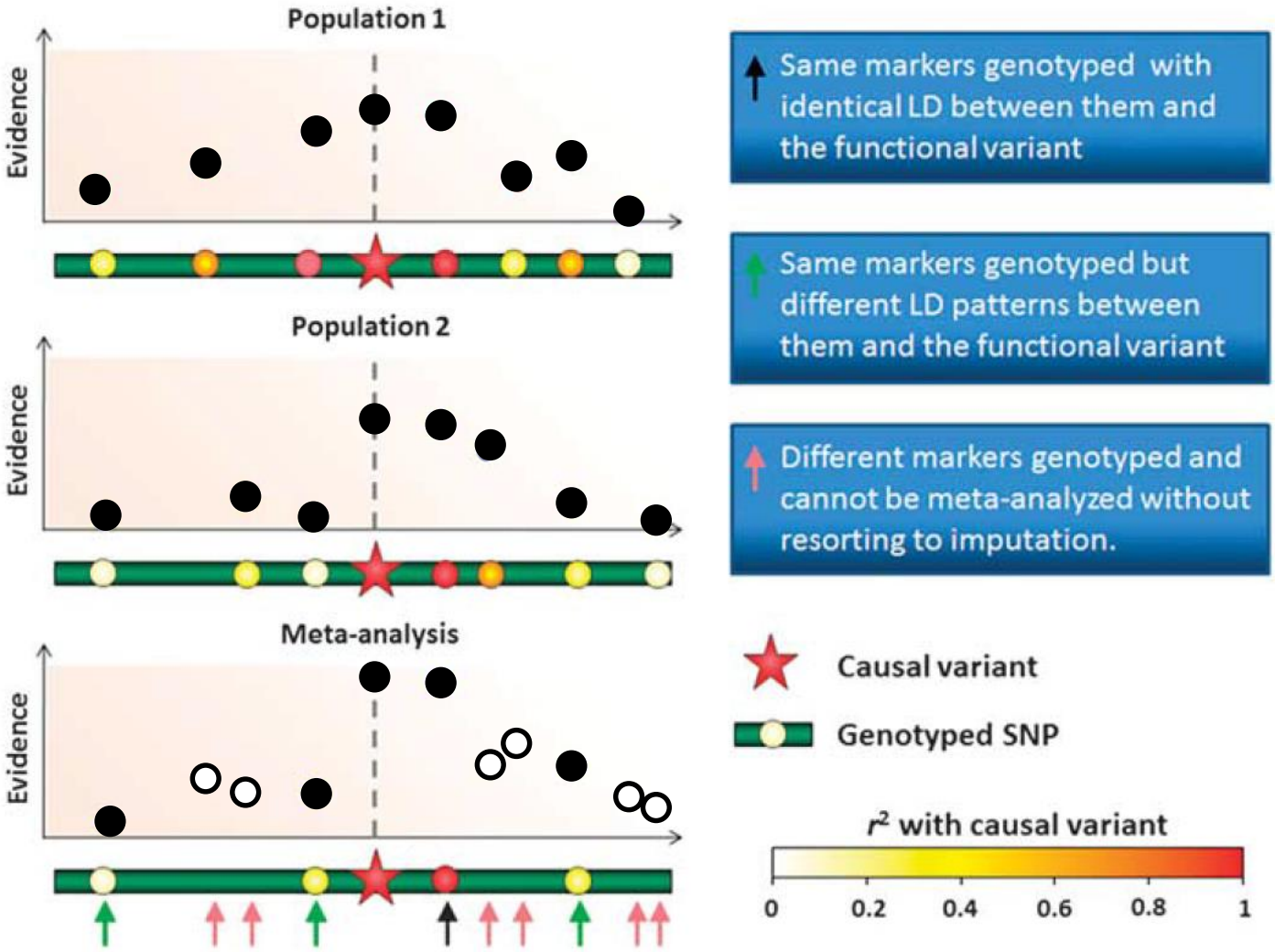
99% credible sets include no more than ten variants at nine T2D susceptibility loci

Locus	Index variant	*p*-value	OR (95% CI)	99% credible set	
				Variants	Interval (bp)
MTNR1B	rs10830963	2.9×10^{-12}	1.10 (1.07-1.13)	1	1
TCF7L2	rs7903146	5.8×10^{-120}	1.39 (1.35-1.43)	3	4,279
KCNQ1	rs74046911	5.9×10^{-18}	1.33 (1.25-1.42)	3	197
ZBED3	rs7732130	6.4×10^{-10}	1.09 (1.06-1.12)	5	10,056
CDKN2A-B	rs10757283	2.8×10^{-19}	1.14 (1.11-1.18)	5	1,007
SLC30A8	rs13266634	1.3×10^{-18}	1.13 (1.10-1.16)	6	33,133
CDKN2A-B	rs10811660	7.0×10^{-43}	1.32 (1.27-1.37)	6	1,397
HNF1B	rs4430796	6.3×10^{-12}	1.09 (1.07-1.12)	7	5,791
CDKAL1	rs35261542	9.6×10^{-23}	1.15 (1.12-1.18)	8	30,073
GLIS3	chr9:4294707:I	6.5×10^{-8}	1.07 (1.05-1.10)	10	15,453

Alternative approaches to fine-mapping causal variants

- Wakefield's approach assumes a single causal variant: hence need for dissection of association signals via (approximate) conditional analysis.
- Recent development of methods that allow for multiple causal variants at a locus: CAVIAR, PAINTOR, FINEMAP, and JAM.
- Methods make use of association summary statistics and reference for LD between variants as the locus.
- Developed in Bayesian framework: MCMC techniques to estimate posterior probability of causality for each variant.
- Can incorporate prior of causality.

Trans-ethnic fine-mapping

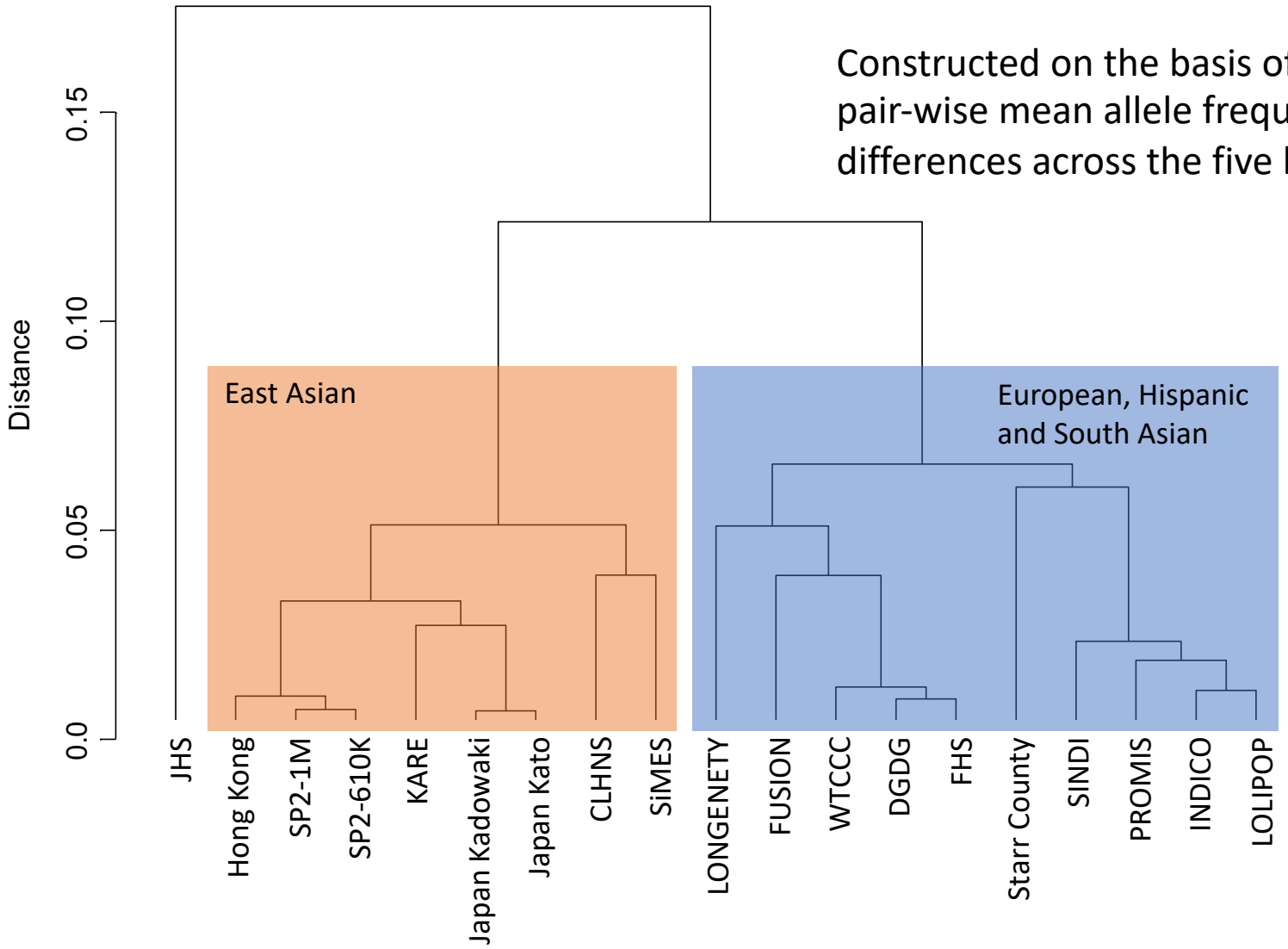


Region-based meta-analysis
X Wang *et al*

Fine-mapping four T2D susceptibility loci

- Meta-analysis of 19 GWAS of 22,086 cases and 42,539 controls from European, South Asian, East Asian, Hispanic and African-American ancestry groups by T2D-GENES Consortium.
- Four T2D loci: *CDKAL1*, *KCNQ1*, *CDKN2A-B*, and *IGF2BP2*:
  - Strongest signals of association in most ethnic groups.
  - Evidence of differences in association signals and patterns of linkage disequilibrium between ethnic groups.
- High-density imputation to 1000 Genomes reference panels provides near complete coverage of common and low-frequency variation.

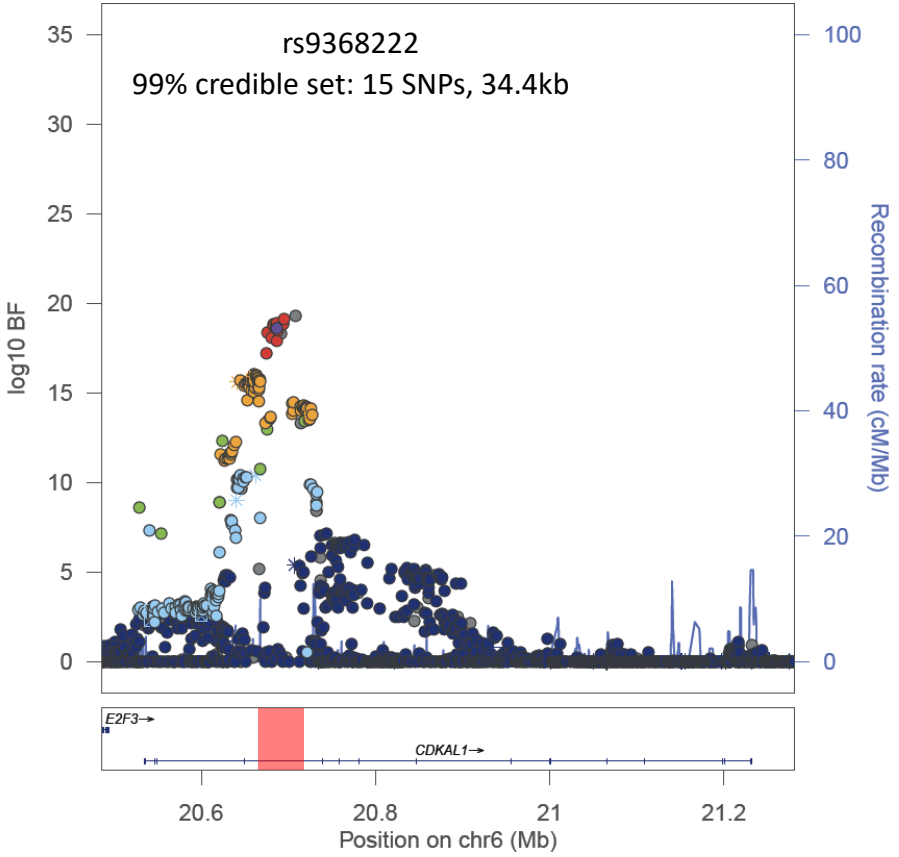
Contributing studies



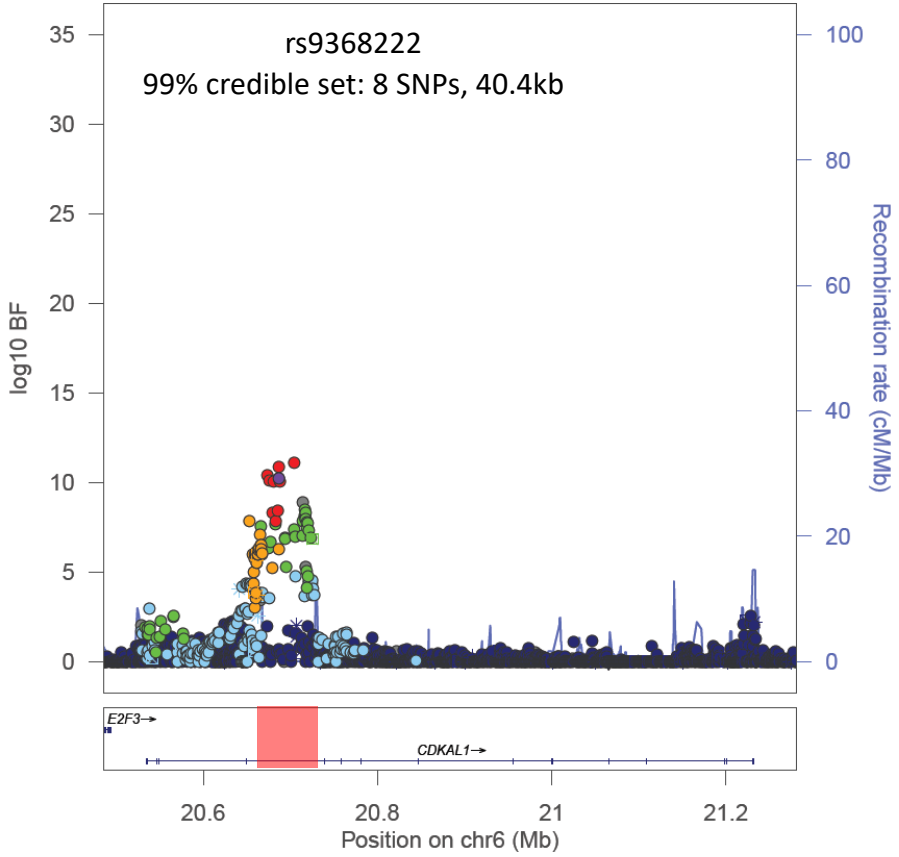
Constructed on the basis of pair-wise mean allele frequency differences across the five loci.

Improved fine-mapping at *CDKAL1*

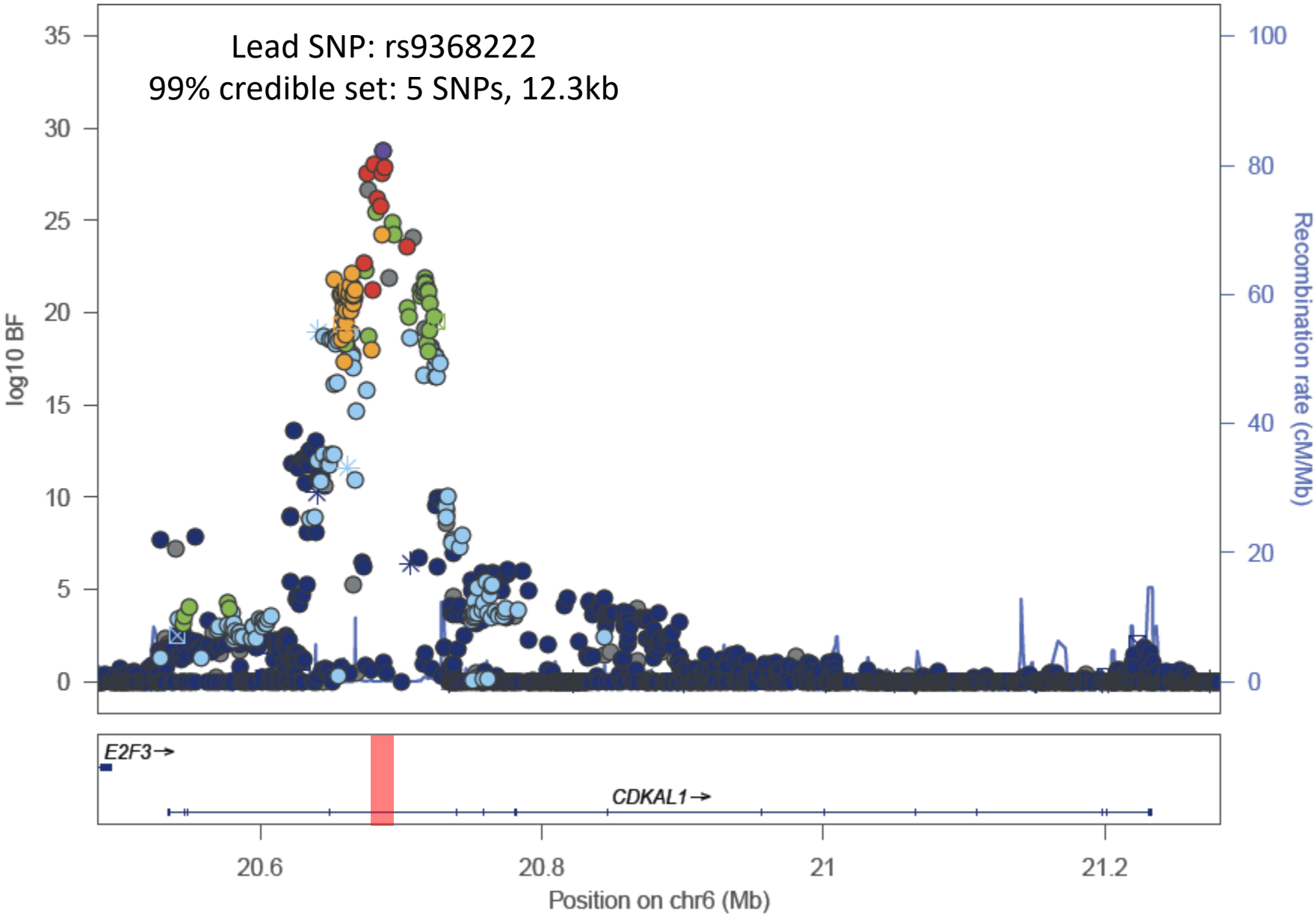
East Asian clade



European, South Asian and Hispanic clade



Improved fine-mapping at *CDKAL1*



A note on trans-ethnic fine-mapping

- Methods that do not assume a single causal variant cannot be directly applied for trans-ethnic fine-mapping.
- Methods require specification of matrix of LD between variants: but LD varies from one ethnic group to another!
- PAINTOR can be used by specifying ethnic-specific association summary statistics and LD matrices.

fGWAS

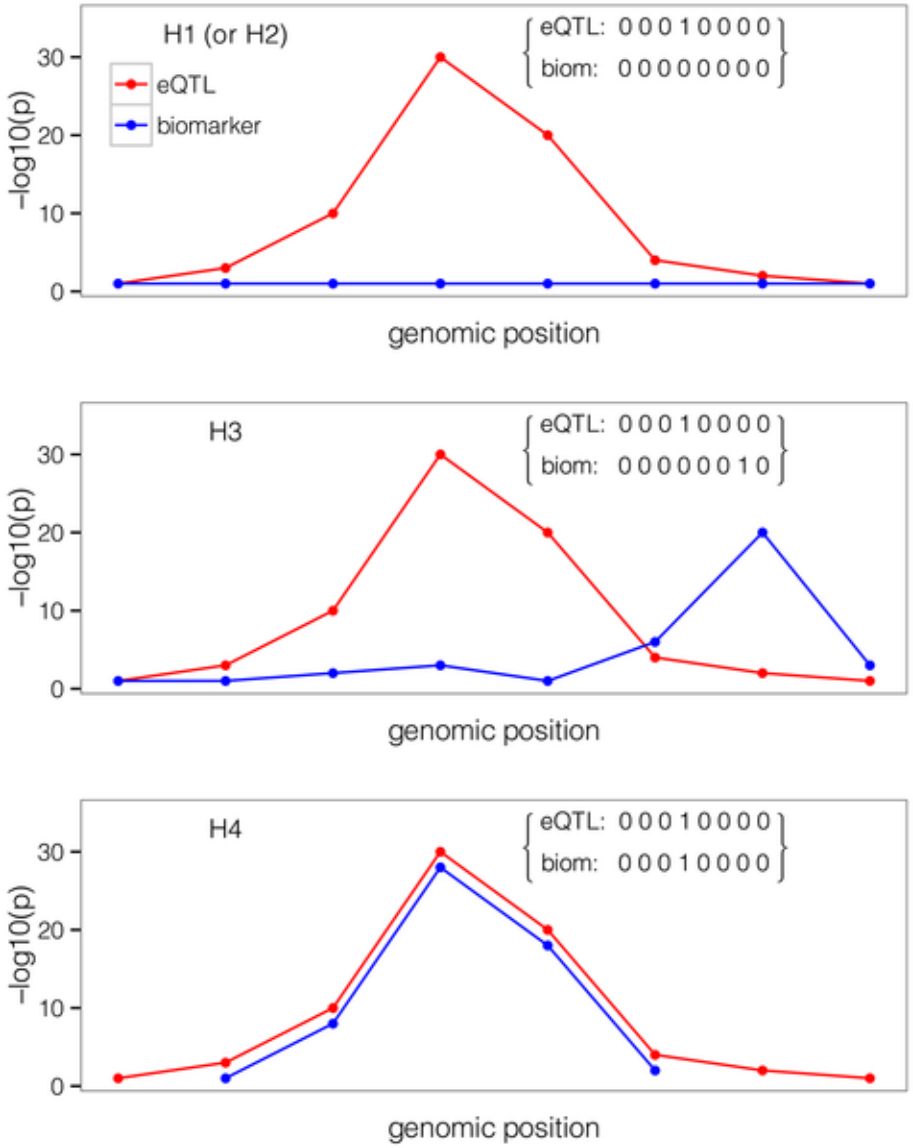
- Evaluates evidence that association signals for a complex trait are enriched in specific genomic annotations to provide insight into upstream biology.
- Protein coding exons.
  - Data available from GENCODE.
- Cis-regulatory elements (CREs): regions of non-coding DNA which regulate the transcription of nearby genes.
  - Typically regulate gene transcription by functioning as binding sites for transcription factors.
  - Most well characterised CREs are enhancers and promoters.
  - May regulate gene transcription in tissue-specific manner.
  - Data available from molecular profiling initiatives such as ENCODE and Epigenome Roadmap Project.

https://github.com/joepickrell/fgwas

Expression quantitative trait loci

- Colocalistion of putative causal variants with expression quantitative trait loci (eQTLs) can provide insight into downstream biology.
- Causal variant impacts trait/disease through regulation of gene expression.
- Regulation of gene expression can be tissue specific or ubiquitous across tissues: important to have trait/disease relevant tissues.
- Data available from GTEx Project (multiple tissues) and GEUVADIS (lymphoblastoid cell lines).

Colocalisation



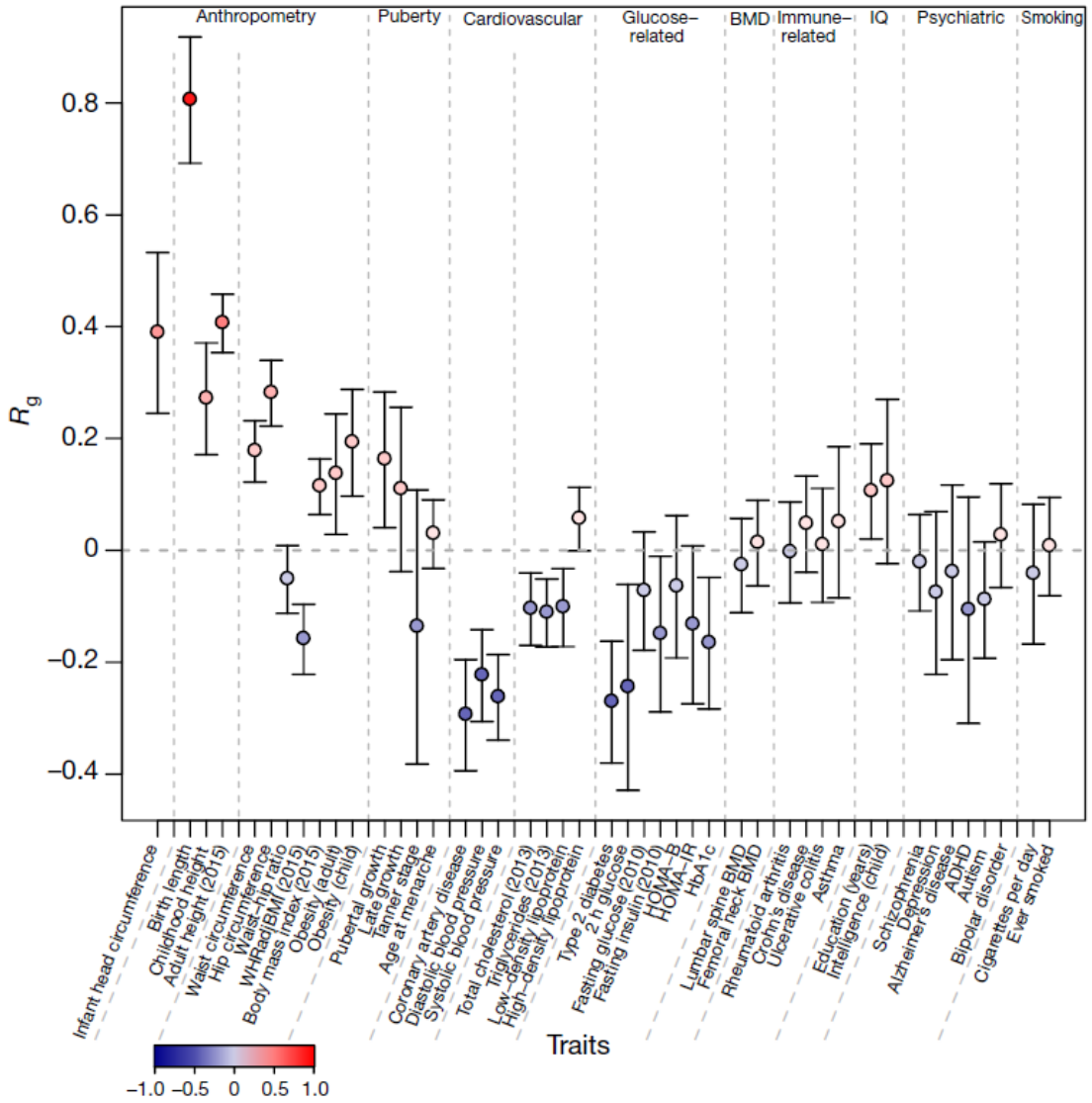
- Compare patterns of association for trait and eQTL.
- Model H1: no association with trait.
- Model H2: no association with eQTL.
- Model H3: association signals with trait and eQTL that are not coincidental.
- Model H4: association signal with trait and eQTL that are coincidental.
- COLOC calculates the posterior probability for each model on the basis of association summary statistics: accounts for LD between variants.

https://github.com/chr1swallace/coloc

LD Score Regression

- LD Score regression uses GWAS summary statistics to estimate SNP heritability of complex traits and diseases genome-wide.
- Important to take account of LD between SNPs: LD score of a variant measures amount of genetic variation tagged by that variant.
- LD scores can be estimated from reference panels, such as 1000 Genomes Project.
- LD score regression can also be used to assess the genetic correlation between pairs of traits based on GWAS summary statistics.
- LD Hub: http://ldsc.broadinstitute.org

Genome-wide associations for birth weight and correlations with adult disease



- Strong positive genetic correlations with anthropometric and obesity-related traits in adults including height ($P=4.8\times10^{-52}$), waist circumference ($P=3.9\times10^{-10}$) and BMI ($P=7.3\times10^{-6}$).
- Strong inverse genetic correlations with indicators of adverse metabolic and cardiovascular health including coronary artery disease ($P=6.5\times10^{-9}$), systolic blood pressure ($P=5.5\times10^{-13}$) and type 2 diabetes ($P=1.1\times10^{-6}$).

Prospects for GWAS

- GWAS will undoubtedly continue to expand the catalogue of regions of the genome contributing to complex human traits!
- Deeply-phenotyped population biobanks with linkage to electronic medical records to evaluate causal relationships between traits (e.g. UK Biobank)
- Development of methods that leverage multi-trait data by modelling the correlation between phenotypes, and offering insight into the shared genetic contribution to human diseases.
- Increasing availability of GWAS in diverse populations, and expanded higher-density reference panels for imputation, such as that from the Haplotype Reference Consortium.
- Improved genomic annotation, particularly in non-coding regions, and expression data from densely genotyped human samples in diverse tissues.
- Development of high-throughput and tractable animal models and relevant in vitro models will allow the functional impact of potential causal genes and variants to be exhaustively assessed.
- Co-ordinated collaboration between researchers over a wide range of disciplines, including human genetics, functional genomics, computational biology and statistical modelling.