Some more advanced techniques used in genome-wide association studies

Luke Jostins-Dean

17/11/2021

GWAS techniques we will cover today

Today we will cover:

- GWAS meta-analysis
- Heritability estimation
- Fine-mapping

We will look at these in the context of two recent(ish) GWAS papers:

- Astle et al (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **17**;167(5):1415-1429.e19 https://pubmed.ncbi.nlm.nih.gov/27863252/

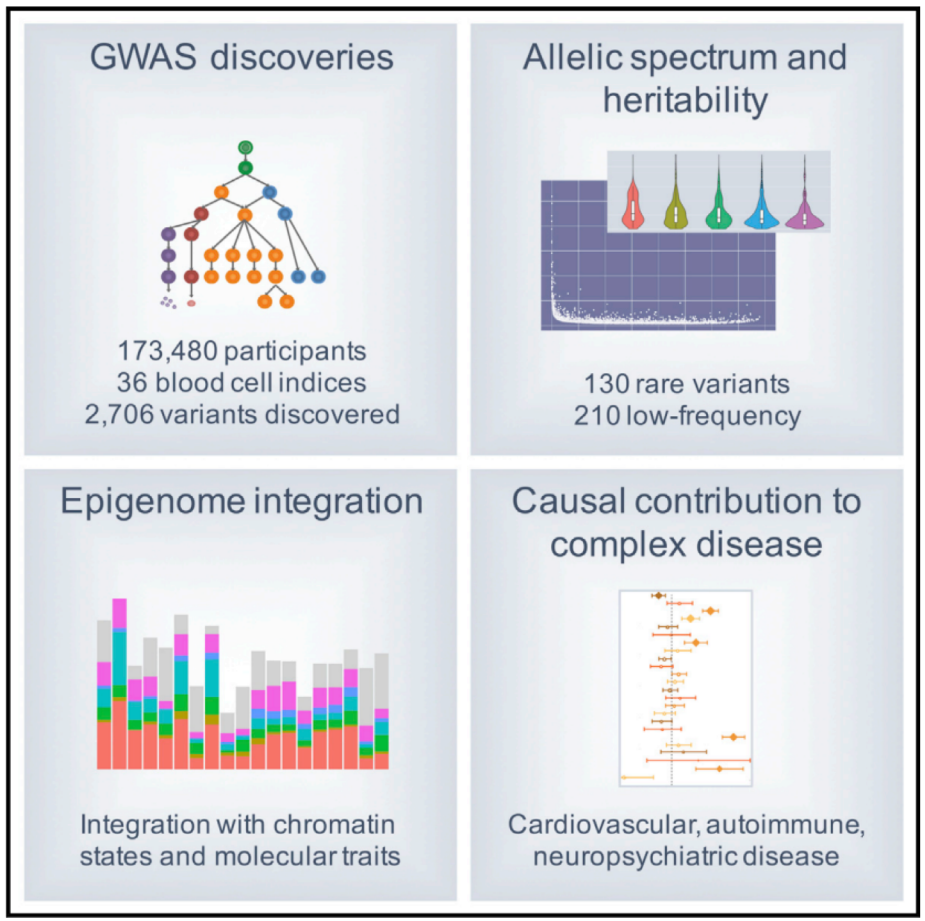
- Robertson, Inshaw, et al (2021) Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nat Genet .* **53**(7):962-971. https://pubmed.ncbi.nlm.nih.gov/34127860/

There will a practical on meta-analysis and fine-mapping, and we will give links to other suggested tutorials and vignettes in the slides.

Cell

The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease

Graphical Abstract



Authors

William J. Astle, Heather Elding, Tao Jiang, ..., Willem H. Ouwehand, Adam S. Butterworth, Nicole Soranzo

Correspondence

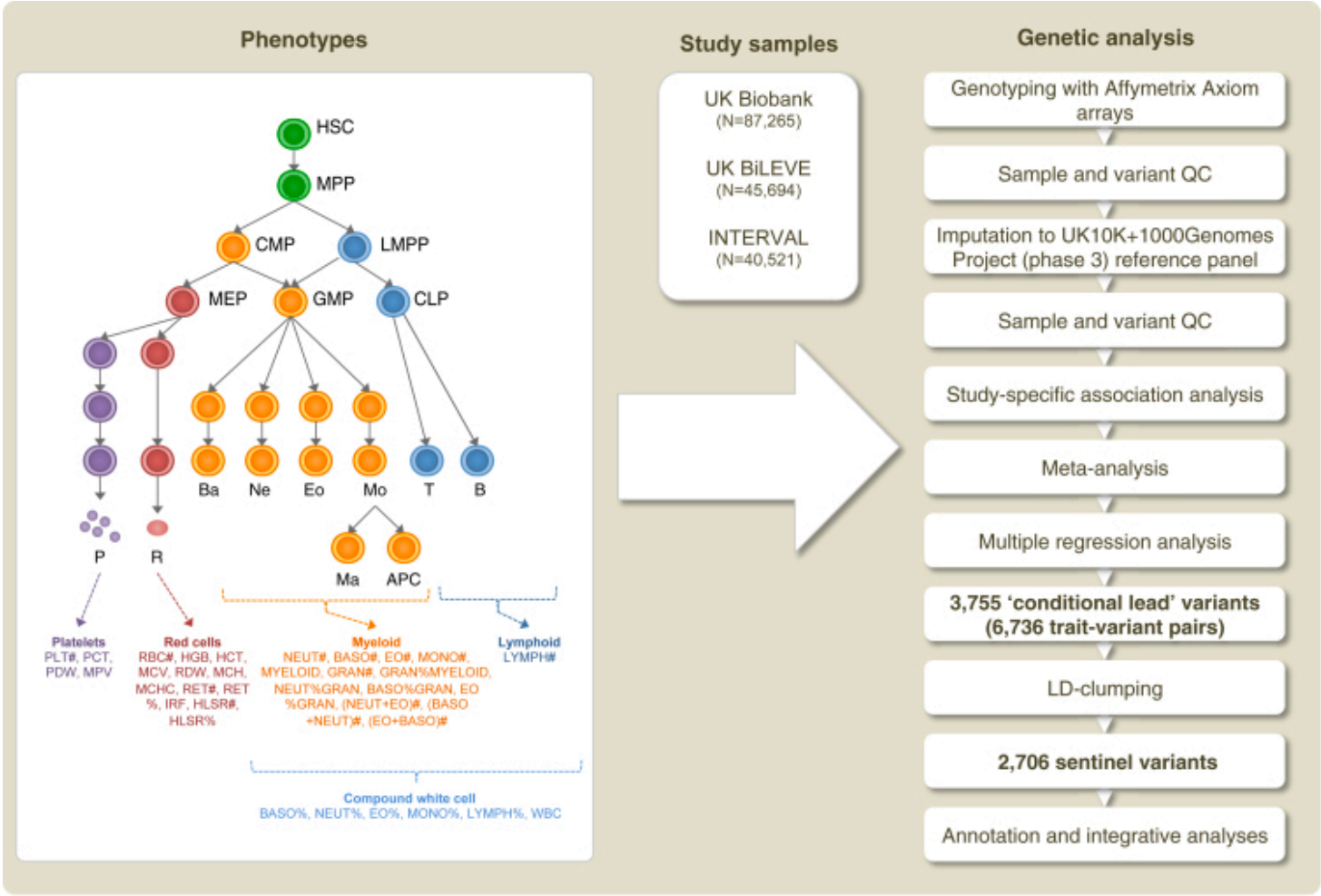
jd292@medschl.cam.ac.uk (J.D.), david.roberts@ndcls.ox.ac.uk (D.J.R.), who1000@cam.ac.uk (W.H.O.), asb38@medschl.cam.ac.uk (A.S.B.), ns6@sanger.ac.uk (N.S.)

In Brief

As part of the IHEC Consortium, this study probes the allelic architecture and regulatory landscape of cellular complex traits with power to identify causal pathways and links to diseases such as schizophrenia. Explore the *Cell Press* IHEC web portal at http://www.cell.com/consortium/IHEC.

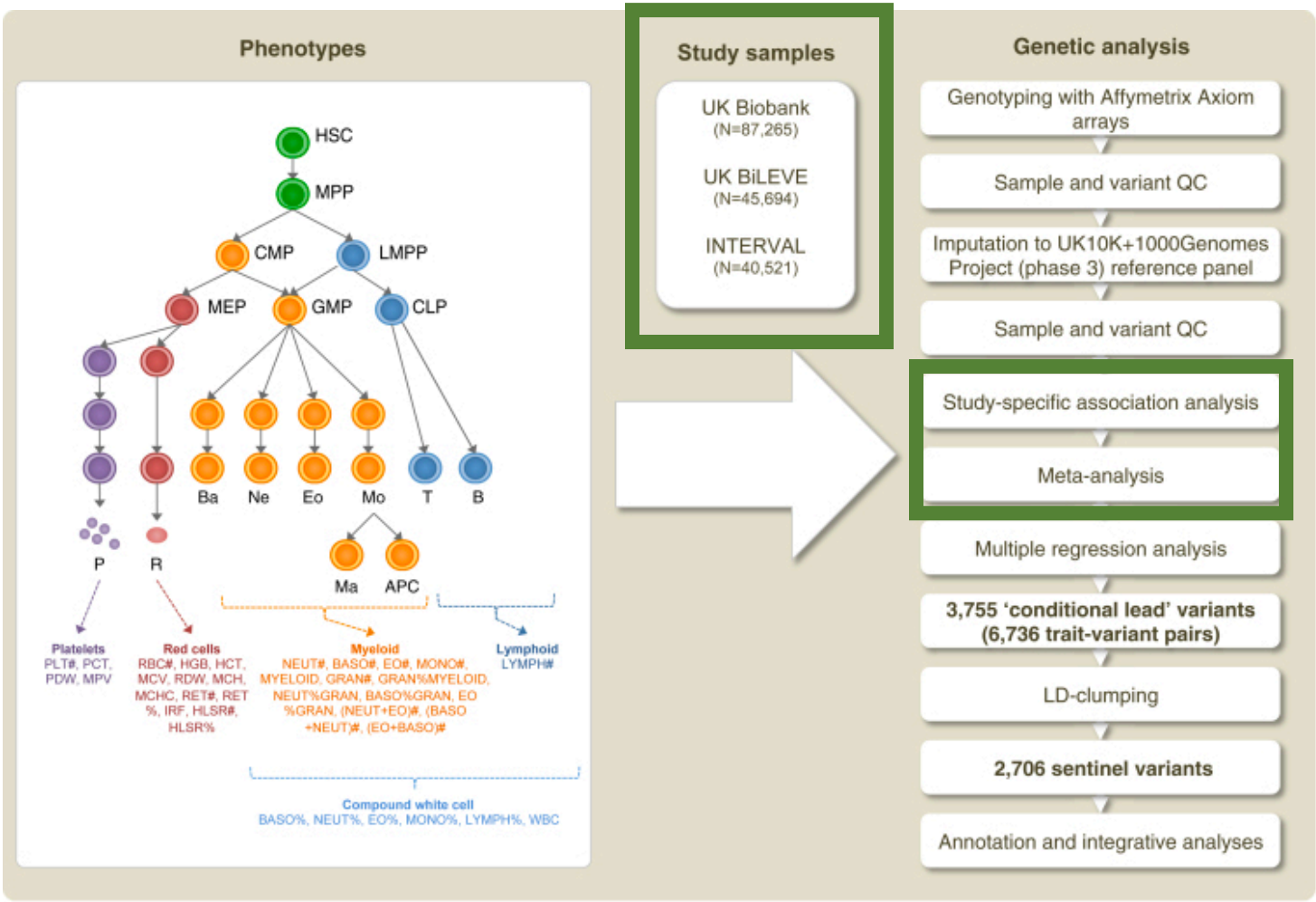
Meta-analysis

Modern large-scale GWAS are usually meta-analyses



HSC = hematopoietic stem cell; MPP = multipotent progenitor; LMPP = lymphomyeloid-restricted progenitors; CMP = common myeloid progenitor; CLP = common lymphoid progenitor; MEP = megakaryocyte and erythroblast progenitor; GMP = granulocyte macrophage progenitor; P = platelet; R = red cell; Ba = basophil; Ne = neutrophil; Eo = eosinophil; Mo = monocyte; Ma = macrophage; APC = antigen presenting cell; T = T-lymphocyte; B = B-lymphocyte.

Modern large-scale GWAS are usually meta-analyses



HSC = hematopoietic stem cell; MPP = multipotent progenitor; LMPP = lymphomyeloid-restricted progenitors; CMP = common myeloid progenitor; CLP = common lymphoid progenitor; MEP = megakaryocyte and erythroblast progenitor; GMP = granulocyte macrophage progenitor; P = platelet; R = red cell; Ba = basophil; Ne = neutrophil; Eo = eosinophil; Mo = monocyte; Ma = macrophage; APC = antigen presenting cell; T = T-lymphocyte; B = B-lymphocyte.

Meta-analyzing genetic data

- Meta-analysis is a technique for combining summary statistics across multiple different studies.
- You need two bits of data for each study, to capture effect size precision
  - e.g. betas and standard errors, or p-values and sample sizes
- The paper uses the software METAL, with settings for Inverse Variance Based analysis
  - This is more properly called a “Fixed Effect Variance Weighted Meta-analysis”

Meta-analyzing genetic data

- Meta-analysis is a technique for combining summary statistics across multiple different studies.
- You need two bits of data for each study, to capture effect size precision
  - e.g. betas and standard errors, or p-values and sample sizes
- The paper uses the software METAL, with settings for Inverse Variance Based analysis
  - This is more properly called a "Fixed Effect Variance Weighted Meta-analysis"

Table 1. Formulae for meta-analysis

	Analytical strategy	
	Sample size based	Inverse variance based
Inputs	N_i - sample size for study i $P_i - P$-value for study i Δ_i - direction of effect for study i	β_i- effect size estimate for study i se_i - standard error for study i
Intermediate Statistics	$Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/\text{SE}_i^2$ $se = \sqrt{1/\sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-Score	$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$	$Z = \beta/SE$
Overall P-value	$P = 2\Phi(\lvert -Z \rvert)$	

Willer et al (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 26(17): 2190–2191.

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score			
P-value			
Weight			

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score			
P-value			
Weight			

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value			
Weight			

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight			

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Sum of weights = 525

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Sum of weights = 525

Meta-analysis effect size = (0.5*100 + 0.2*400 + 0.4*25)/(525) = 0.267

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Sum of weights = 525

Meta-analysis effect size = (0.5*100 + 0.2*400 + 0.4*25)/(525) = 0.267

Meta-analysis standard error = sqrt(1/525) = 0.0436

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

Sum of weights = 525

Meta-analysis effect size = (0.5*100 + 0.2*400 + 0.4*25)/(525) = 0.267
Meta-analysis standard error = sqrt(1/525) = 0.0436
Meta-analysis Z = 0.267/0.0436 = 6.12

Meta-analyzing two studies with variance-weighted fixed-effect meta-analysis

	Study 1	Study 2	Study 3
Effect size	0.5	0.2	0.4
Standard error	0.1	0.05	0.2
Z score	0.5/0.1 = 5	0.2/0.05=4	0.4/0.2=2
P-value	2*pnorm(-5) = 5.7e-7	2*pnorm(-4)=6.3e-5	2*pnorm(-2)=0.046
Weight	1/0.1^2 = 100	1/0.05^2=400	1/0.2^2 = 25

Inverse variance based

β_i- effect size estimate for study i

se_i - standard error for study i

$w_i = 1/\mathrm{SE}_i^2$

$se = \sqrt{1/\sum_i w_i}$

$\beta = \sum_i \beta_i w_i / \sum_i w_i$

$Z = \beta / SE$

$P = 2\Phi(|-Z|)$

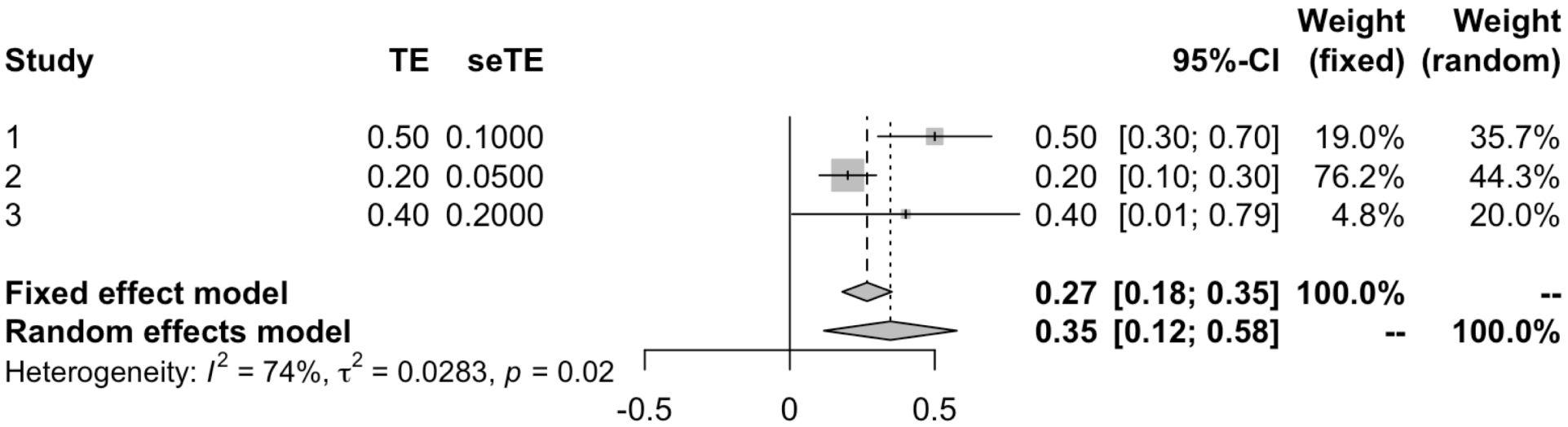
Sum of weights = 525

Meta-analysis effect size = (0.5*100 + 0.2*400 + 0.4*25)/(525) = 0.267
Meta-analysis standard error = sqrt(1/525) = 0.0436
Meta-analysis Z = 0.267/0.0436 = 6.12
Meta-analysis p = 9.4e-10

Some things to be keep in mind

- Fixed effect meta-analysis assumes that them true effect size is exactly the same in each study.
  - It also requires them to be on the same SCALE
  - E.g. meta-analyzing a study measured in kg and one measured in lbs would not work as expected – you first need to convert them onto the same scale.
- You can test whether this assumption is violated using a heterogeneity test, e.g. the Cochran Q test (implement in most meta-analysis software).
- There are lots of other options for meta-analysis:
  - Random effects meta-analysis: effect size assumed to vary across studies, normally distributed with some variance tau^2
  - Trans-ethnic meta-analysis: a specific type of random effect meta-analysis designed to study genetics across populations (the MANTRA software is an example).
  - Meta-regression: designed to answer the question “why do these studies differ?”, by including per-study covariates (eg average age, ethnicity, etc).

Visualizing meta-analysis results: Forest plot

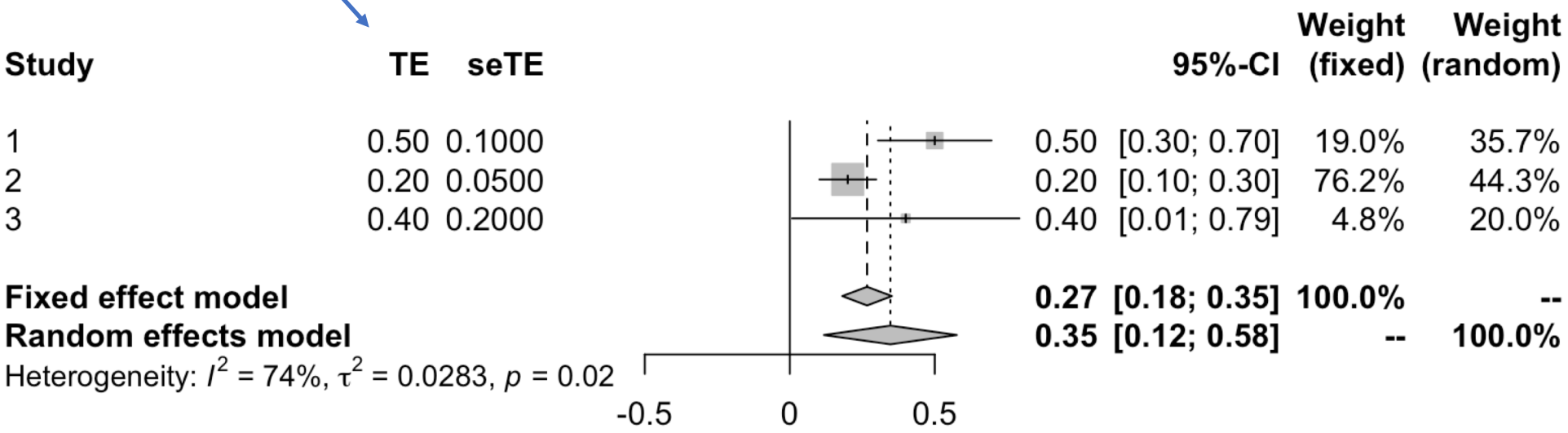


Default forest plotting using R package “meta”:

> forest.meta(metagen(betas,ses))

Visualizing meta-analysis results: Forest plot

TE="Treatment effect", i.e. effect size and its standard error



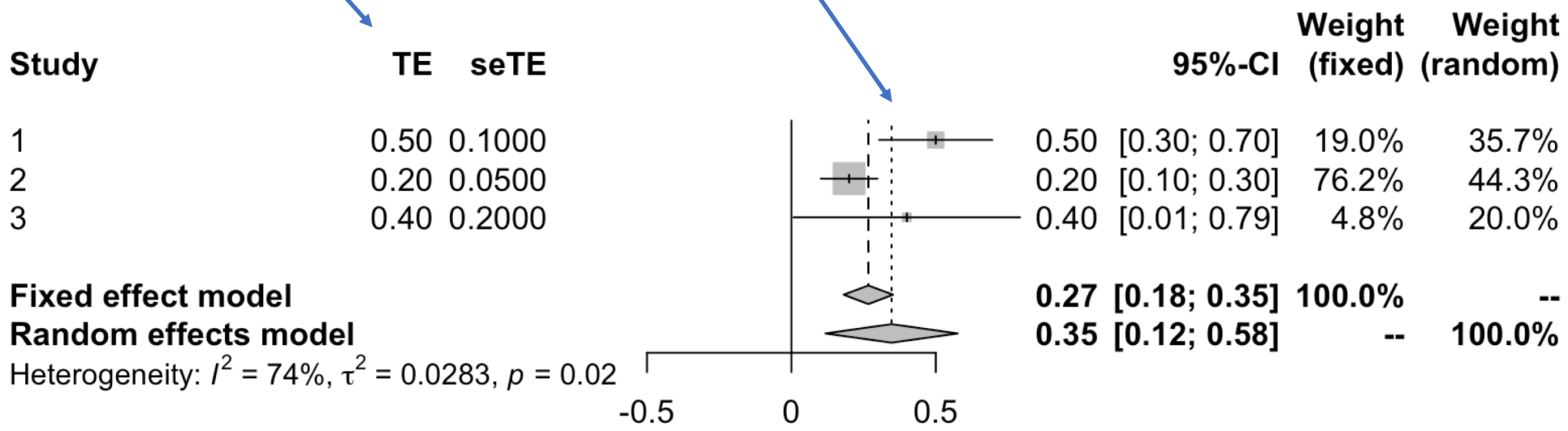
Default forest plotting using R package "meta":

> forest.meta(metagen(betas,ses))

Visualizing meta-analysis results: Forest plot

TE="Treatment effect", i.e. effect size and its standard error

Confidence intervals on effect size for each individual study



Default forest plotting using R package "meta":

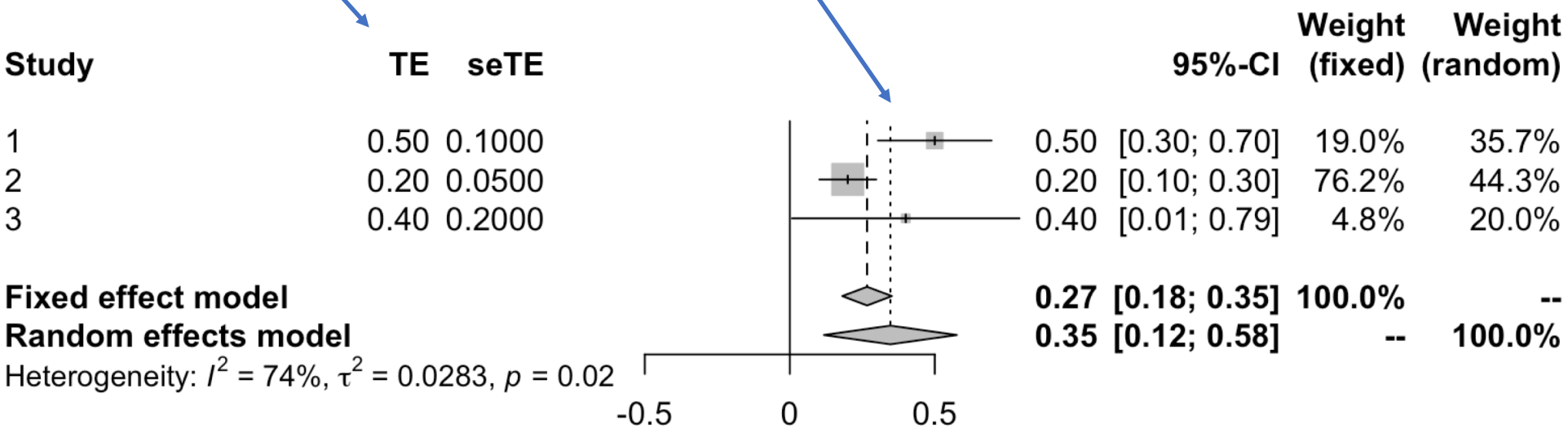
> forest.meta(metagen(betas,ses))

Visualizing meta-analysis results: Forest plot

TE="Treatment effect", i.e. effect size and its standard error

Confidence intervals on effect size for each individual study

Weights:
1/se^2 for fixed effect
1/(se^2 + tau^2) for random effects



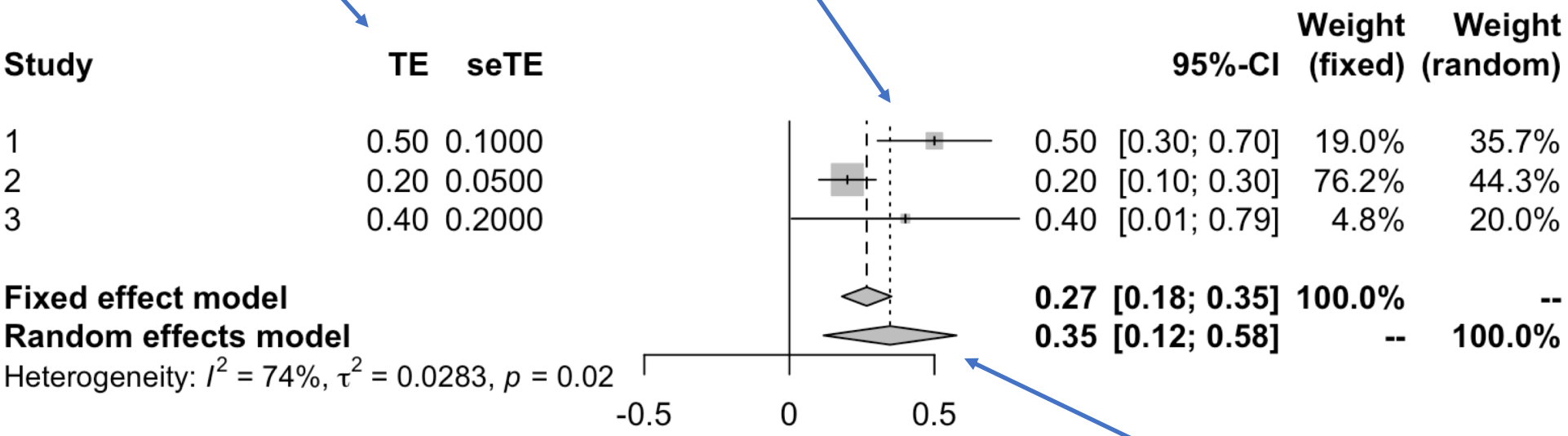
Default forest plotting using R package "meta":
> forest.meta(metagen(betas,ses))

Visualizing meta-analysis results: Forest plot

TE="Treatment effect", i.e. effect size and its standard error

Confidence intervals on effect size for each individual study

Weights:
1/se^2 for fixed effect
1/(se^2 + tau^2) for random effects



Confidence intervals for meta-analysed effect size

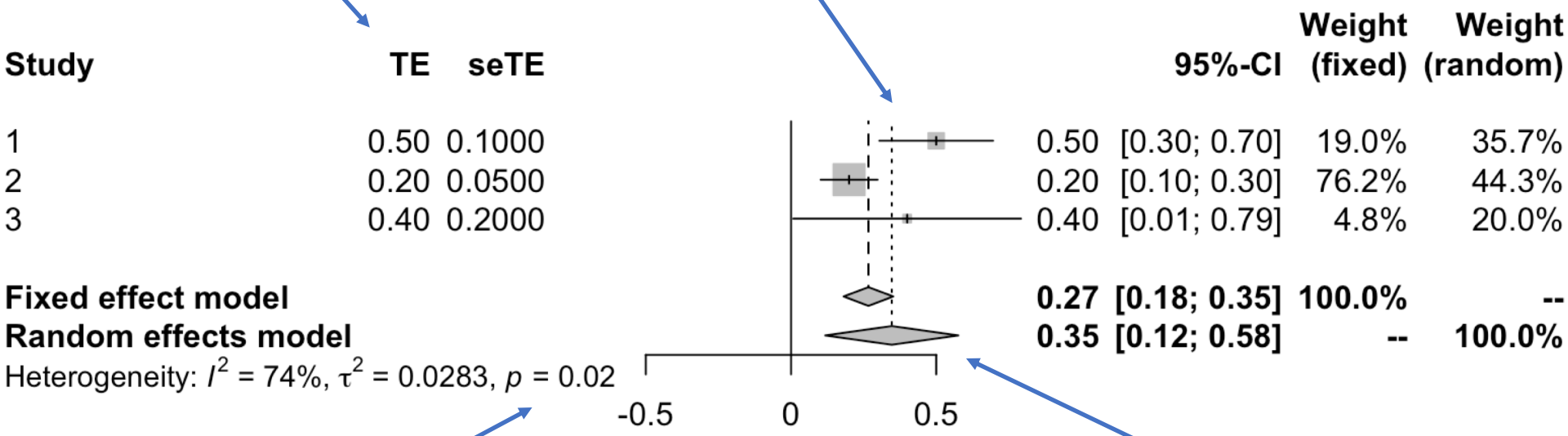
Default forest plotting using R package "meta":
> forest.meta(metagen(betas,ses))

Visualizing meta-analysis results: Forest plot

TE="Treatment effect", i.e. effect size and its standard error

Confidence intervals on effect size for each individual study

Weights:
1/se^2 for fixed effect
1/(se^2 + tau^2) for random effects

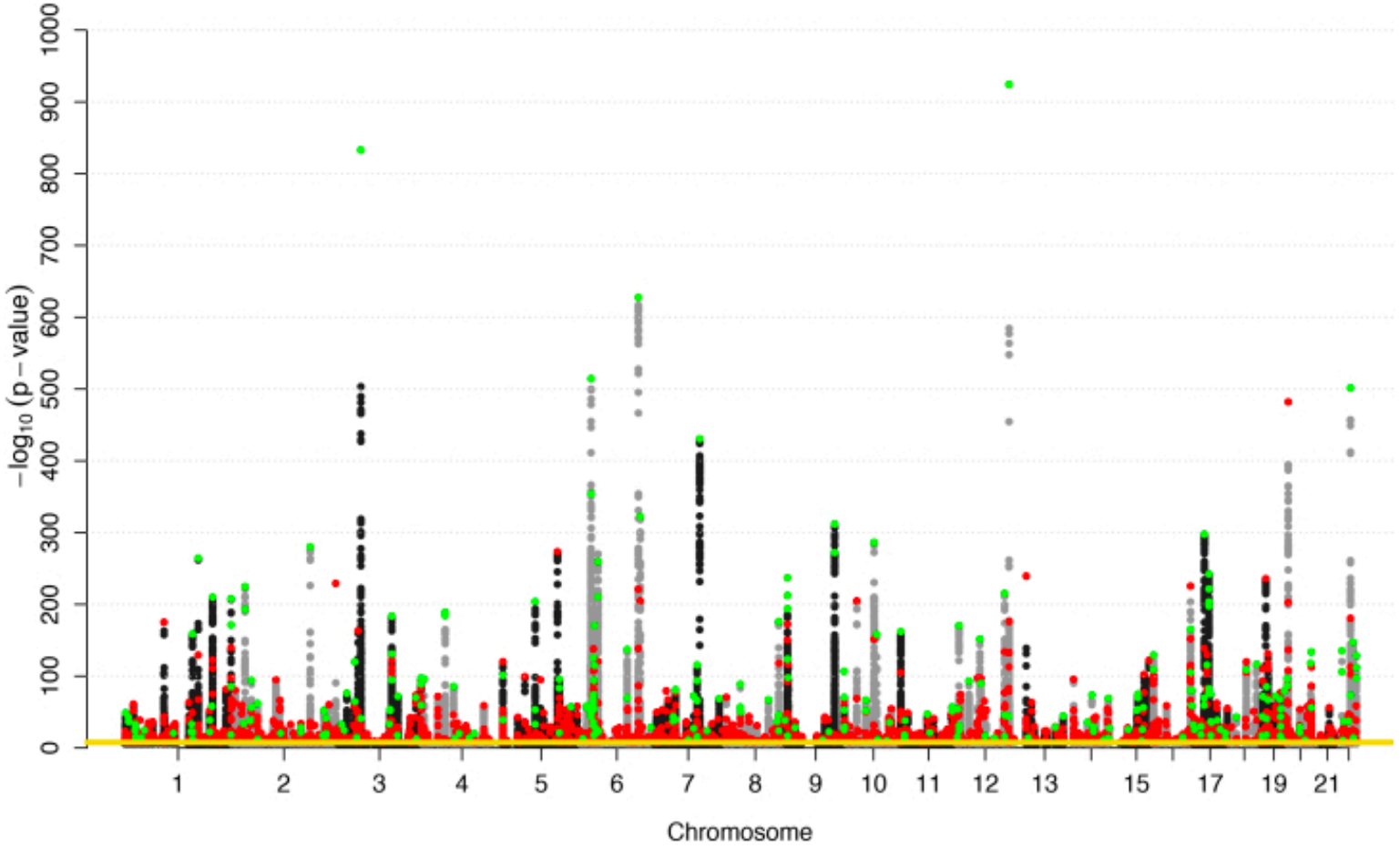


Measures of study-to-study variation.

Confidence intervals for meta-analysed effect size

Default forest plotting using R package "meta":
> forest.meta(metagen(betas,ses))

Back to the paper



These are the results of the meta-analysis for every SNP and every blood trait. Red dots are new top hits, green are known top hits, and black are variants that are in LD with either.

The authors were worried about heterogeneity

Heterogeneity Filtering Substantial statistical evidence for heterogeneity in effect sizes between the studies of a meta-analysis for a genome-wide significant variant is often taken to suggest a false-positive association. However, effect size heterogeneity in GWAS can be generated by:

- population-genotype interactions (i.e., true allelic effect size differences between studies),
- variation in LD between study populations,
- study specific quantile-inverse-normal transformations, when there are differences in the adjustment of phenotypes for covariates between studies,
- differences in genotyping measurement error between studies (when independent of phenotype, such errors tend to bias associations toward the null) and
- differences in phenotyping measurement techniques between studies, none of which are necessarily reasons to regard an observed population association as spurious.

The authors were worried about heterogeneity

Heterogeneity Filtering Substantial statistical evidence for heterogeneity in effect sizes between the studies of a meta-analysis for a genome-wide significant variant is often taken to suggest a false-positive association. However, effect size heterogeneity in GWAS can be generated by:

- population-genotype interactions (i.e., true allelic effect size differences between studies),
- variation in LD between study populations,
- study specific quantile-inverse-normal transformations, when there are differences in the adjustment of phenotypes for covariates between studies,
- differences in genotyping measurement error between studies (when independent of phenotype, such errors tend to bias associations toward the null) and
- differences in phenotyping measurement techniques between studies, none of which are necessarily reasons to regard an observed population association as spurious.

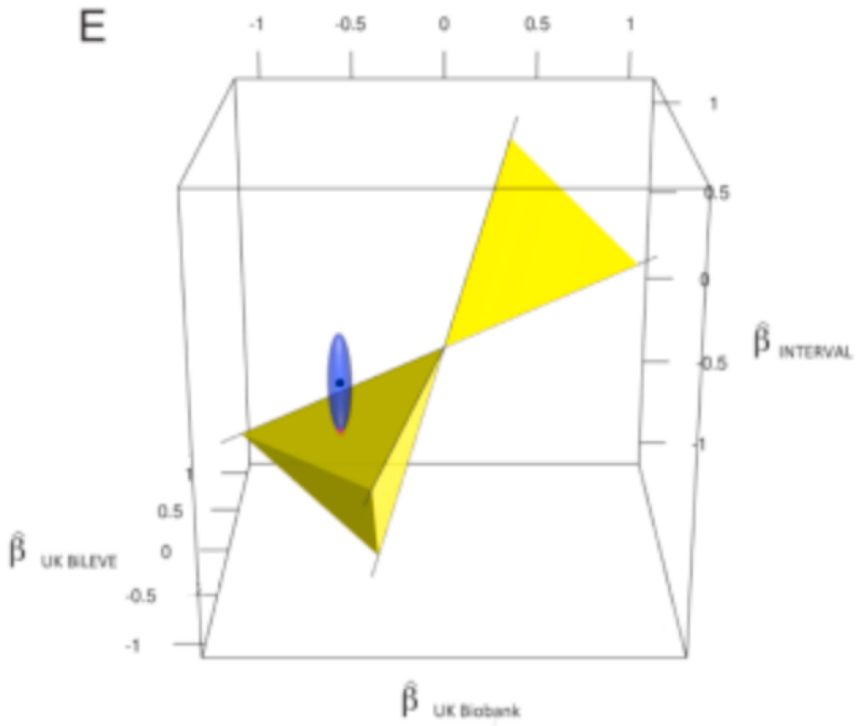
Due to the high power of the present analysis, we found that common variants showing directionally concordant evidence for association across the three studies were often removed when we filtered variants by thresholding a statistic measuring evidence for quantitative heterogeneity in effect size (Cochran's Q). Consequently, we devised an alternative (generalized) statistic to detect heterogeneity in effect size that we regard as implausible for genuine population associations. The three dimensional plot (Figure S2E) illustrates our approach.

The authors were worried about heterogeneity

Heterogeneity Filtering Substantial statistical evidence for heterogeneity in effect sizes between the studies of a meta-analysis for a genome-wide significant variant is often taken to suggest a false-positive association. However, effect size heterogeneity in GWAS can be generated by:

- population-genotype interactions (i.e., true allelic effect size differences between studies),
- variation in LD between study populations,
- study specific quantile-inverse-normal transformations, when there are differences in the adjustment of phenotypes for covariates between studies,
- differences in genotyping measurement error between studies (when independent of phenotype, such errors tend to bias associations toward the null) and
- differences in phenotyping measurement techniques between studies, none of which are necessarily reasons to regard an observed population association as spurious.

Due to the high power of the present analysis, we found that common variants showing directionally concordant evidence for association across the three studies were often removed when we filtered variants by thresholding a statistic measuring evidence for quantitative heterogeneity in effect size (Cochran's Q). Consequently, we devised an alternative (generalized) statistic to detect heterogeneity in effect size that we regard as implausible for genuine population associations. The three dimensional plot (Figure S2E) illustrates our approach.



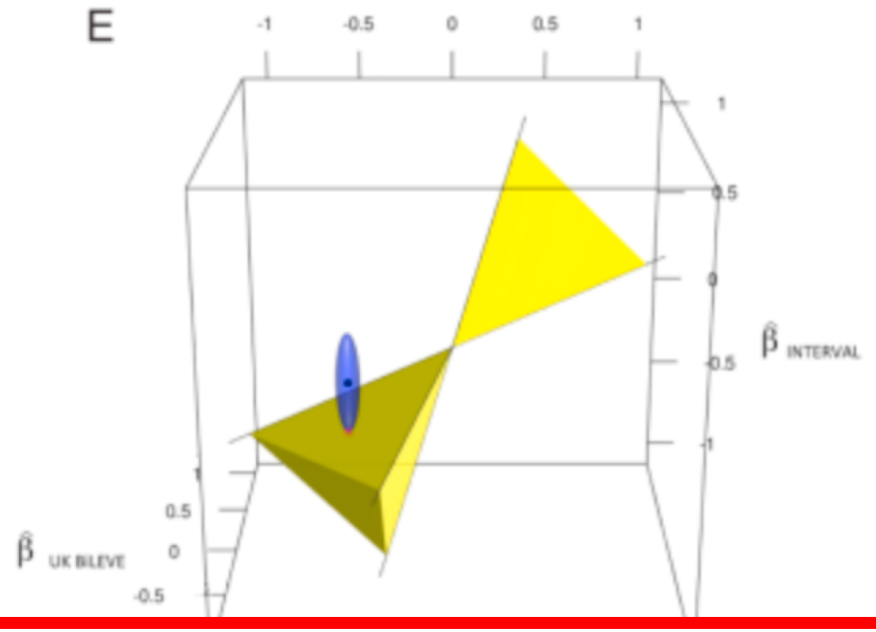
(E) Illustration of the method used to determine the weight of evidence that heterogeneity in effect sizes across the three studies exceeded a tolerance criterion. The axes represent effect sizes in UK Biobank, INTERVAL and UK BiLEVE. The black dot represents the vector of study specific effect size estimates ($\hat{\beta}_{\text{UK Biobank}}, \hat{\beta}_{\text{INTERVAL}}, \hat{\beta}_{\text{UK BiLEVE}}$) for a variant. If the dot lies inside the infinite yellow double-pyramid (defined by three planes intersecting the origin, each normal to one of $n_1 = (1, -1/4, -1/4)$, $n_2 = (-1/4, 1, -1/4)$, $n_3 = (-1/4, -1/4, 1)$) we consider that there is no evidence of between study heterogeneity. If the black dot lies outside the yellow double-pyramid we measure the strength of evidence for heterogeneity as the distance between the black dot and the nearest point on the surface of the pyramid (red dot), with distances scaled to account for the standard errors of the study specific estimators. The nearest point on the pyramid is thus defined as the point in the smallest confidence surface for the estimators that intersects the pyramid (blue ellipsoid). We thresholded the distance score at 5.2 and filtered all variant-blood index pairs exceeding the score from further analysis.

The authors were worried about heterogeneity

Heterogeneity Filtering Substantial statistical evidence for heterogeneity in effect sizes between the studies of a meta-analysis for a genome-wide significant variant is often taken to suggest a false-positive association. However, effect size heterogeneity in GWAS can be generated by:

- population-genotype interactions (i.e., true allelic effect size differences between studies),
- variation in LD between study populations,
- study specific quantile-inverse-normal transformations, when there are differences in the adjustment of phenotypes for covariates between studies,
- differences in genotyping measurement error between studies (when independent of phenotype, such errors tend to bias associations toward the null) and
- differences in phenotyping measurement techniques between studies, none of which are necessarily reasons to regard an observed population association as spurious.

Due to the high power of the present analysis, we found that common variants showing directionally concordant evidence for association across the three studies were often removed when we filtered variants by thresholding a statistic measuring evidence for quantitative heterogeneity in effect size (Cochran's Q). Consequently, we devised an alternative (generalized) statistic to detect heterogeneity in effect size that we regard as implausible for genuine population associations. The three dimensional plot (Figure S2E) illustrates our approach.



TL;DR: The authors tolerate some amount of between-study heterogeneity, providing it lies within plausible bounds.

(E) Illustrat… across the th… INTERVAL… UK Biobank,… (defined by… −1/4), $n_3 = (-1/4,-1/4, 1)$) we consider that there is no evidence of between study heterogeneity. If the black dot lies outside the yellow double-pyramid we measure the strength of evidence for heterogeneity as the distance between the black dot and the nearest point on the surface of the pyramid (red dot), with distances scaled to account for the standard errors of the study specific estimators. The nearest point on the pyramid is thus defined as the point in the smallest confidence surface for the estimators that intersects the pyramid (blue ellipsoid). We thresholded the distance score at 5.2 and filtered all variant-blood index pairs exceeding the score from further analysis.

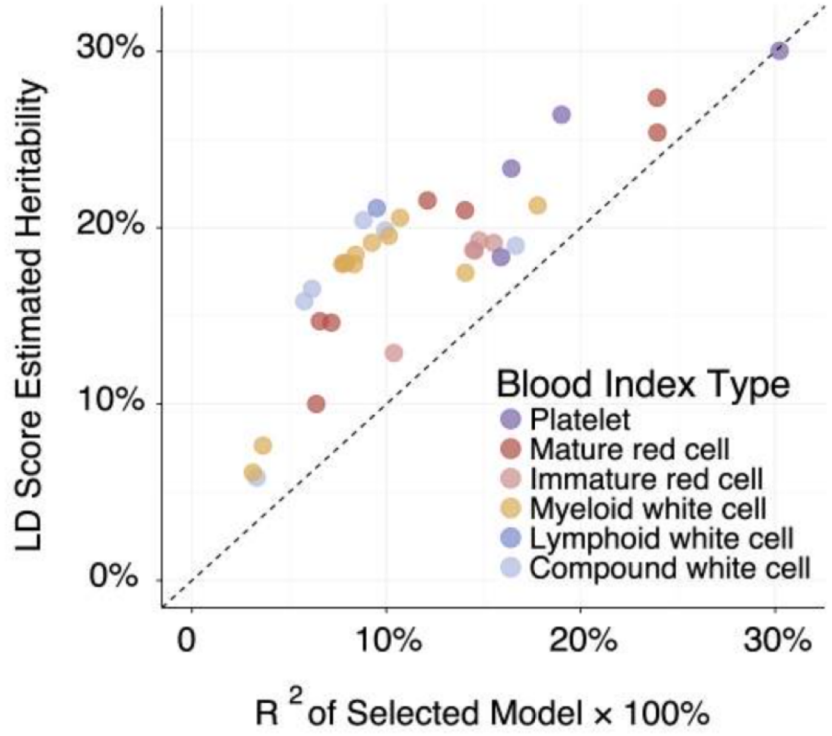
Running these yourself

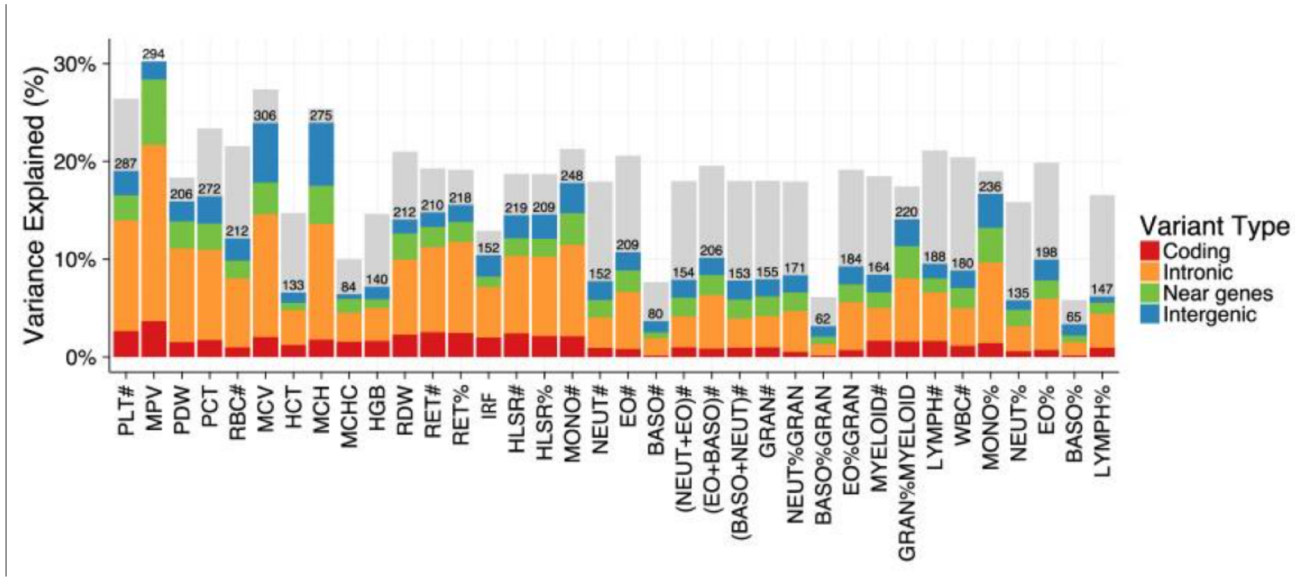
- The practical today will cover meta-analysed genome-wide summary statistics from two studies, and visualizing them using forest plots.
- The METAL paper (cited earlier in the talk) is also quite short and readable, and they also have a Quick Start tutorial that includes example data:

  https://genome.sph.umich.edu/wiki/METAL_Quick_Start

Heritability estimation

A lot of figures involve heritability, variance explained, R2, etc





The broad-sense heritability

$$y = \hat{y}(G) + e$$

The broad-sense heritability

$$y = \hat{y}(G) + e$$

Actual phenotype

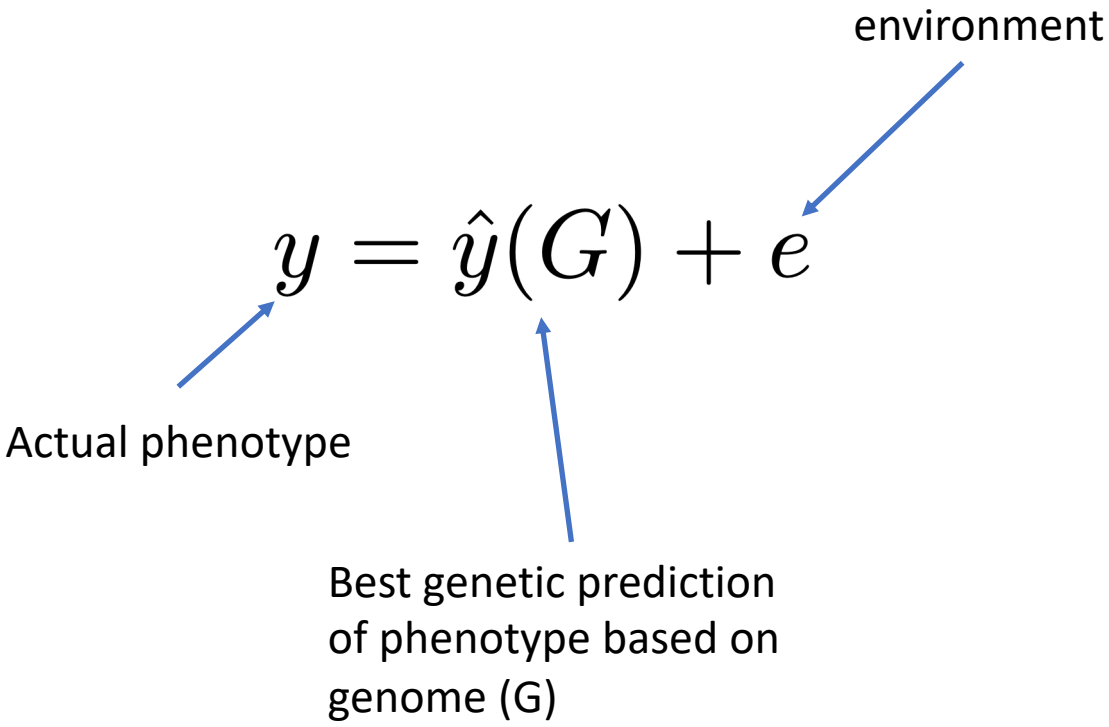
The broad-sense heritability

$$y = \hat{y}(G) + e$$

Actual phenotype

Best genetic prediction of phenotype based on genome (G)

The broad-sense heritability



The broad-sense heritability

environment

$$y = \hat{y}(G) + e$$

Actual phenotype

Best genetic prediction of phenotype based on genome (G)

$$y \sim N(0, 1)$$

$$\hat{y}(G) \sim N(0, H^2)$$

$$e \sim N(0, 1 - H^2)$$

The broad-sense heritability

environment

$$y = \hat{y}(G) + e$$

Actual phenotype

Best genetic prediction of phenotype based on genome (G)

$$y \sim N(0, 1)$$

$$\hat{y}(G) \sim N(0, H^2)$$

$$e \sim N(0, 1 - H^2)$$

The squared correlation between the best possible genetic predictor and the real phenotype is the broad sense heritability, H2:

$$cor(y, \hat{y}(G))^2 = H^2$$

The broad-sense heritability

environment

$$y = \hat{y}(G) + e$$

Actual phenotype

Best genetic prediction of phenotype based on genome (G)

$$y \sim N(0, 1)$$

$$\hat{y}(G) \sim N(0, H^2)$$

$$e \sim N(0, 1 - H^2)$$

The squared correlation between the best possible genetic predictor and the real phenotype is the broad sense heritability, H2:

$$cor(y, \hat{y}(G))^2 = H^2$$

The broad sense heritability is a measure of the totality of genetic effects. It is mostly theoretical, though it is (in theory) equal to the correlation in phenotype of identical twins raised apart.

The narrow-sense heritability

$$y = \beta^T G + e$$

$$y \sim N(0, 1)$$

$$\beta^T G \sim N(0, h^2)$$

$$e \sim N(0, 1 - h^2)$$

The squared correlation between the best possible linear genetic predictor and the real phenotype is the narrow sense heritability, h2:

$$cor(y, \beta^T G)^2 = h^2$$

The narrow-sense heritability

Environment plus non-linear genetics

$$y = \beta^T G + e$$

Actual phenotype

Best LINEAR genetic prediction of phenotype based on genome (G)

$$y \sim N(0, 1)$$

$$\beta^T G \sim N(0, h^2)$$

$$e \sim N(0, 1 - h^2)$$

The squared correlation between the best possible linear genetic predictor and the real phenotype is the narrow sense heritability, h2:

$$cor(y, \beta^T G)^2 = h^2$$

The narrow-sense heritability

Environment plus non-linear genetics

$$y = \beta^T G + e$$

Actual phenotype

Best LINEAR genetic prediction of phenotype based on genome (G)

$$y \sim N(0, 1)$$

$$\beta^T G \sim N(0, h^2)$$

$$e \sim N(0, 1 - h^2)$$

The squared correlation between the best possible linear genetic predictor and the real phenotype is the narrow sense heritability, h2:

$$cor(y, \beta^T G)^2 = h^2$$

The narrow sense heritability is the extent to which genetics 'breeds true', i.e. is passed down in families i.e. where the correlation in phenotype is proportional to relatedness.

The SNP-heritability

Environment, non-linear genetics, snps not in the study

$$y = \beta_{snp}^T G_{snp} + e$$

Actual phenotype

Best linear genetic prediction of phenotype based on the set of snps that are in the study (G_snp)

$$y \sim N(0, 1)$$

$$\beta_{snp}^T G_{snp} \sim N(0, h^2)$$

$$e \sim N(0, 1 - h^2)$$

The squared correlation between the best possible linear genetic predictor using the snps in your study and the real phenotype is the NNP heritability, h2_snp:

$$cor(y, \beta_{snp}^T G_{snp})^2 = h_{snp}^2$$

The SNP-heritability

Environment, non-linear genetics, snps not in the study

$$y = \beta_{snp}^T G_{snp} + e$$

Actual phenotype

Best linear genetic prediction of phenotype based on the set of snps that are in the study (G_snp)

$$y \sim N(0,1)$$

$$\beta_{snp}^T G_{snp} \sim N(0, h^2)$$

$$e \sim N(0, 1-h^2)$$

The SNP heritability measured the total narrow-sense heritability captured, i.e. the "variance explained", by the variants you have studied. In a GWAS, this usually means "captured by common variants". This gives a lower bound for the narrow-sense heritability.

The squared correlation between the best possible linear genetic predictor using the snps in your study and the real phenotype is the NNP heritability, h2_snp:

$$cor(y, \beta_{snp}^T G_{snp})^2 = h_{snp}^2$$

Estimating the SNP heritability using a polygenic risk score

Estimate a polygenic risk score, by trying to estimate the effect sizes (beta^hat):

Estimating the SNP heritability using a polygenic risk score

Estimate a polygenic risk score, by trying to estimate the effect sizes (beta^hat):

$$\hat{\beta}_{snp}^{T} G_{snp}$$

Estimated weights for each snp

Estimating the SNP heritability using a polygenic risk score

Estimate a polygenic risk score, by trying to estimate the effect sizes (beta^hat):

$$\hat{\beta}_{snp}^{T} G_{snp}$$

Estimated weights for each snp

Crude method: estimate betas for genome-wide significant hits and set beta = 0 for everything else.
More sophisticated methods: use a lasso or a Bayesian prior to shrink effect sizes genome-wide.

Estimating the SNP heritability using a polygenic risk score

Estimate a polygenic risk score, by trying to estimate the effect sizes (beta^hat):

$$\hat{\beta}_{snp}^{T} G_{snp}$$

Crude method: estimate betas for genome-wide significant hits and set beta = 0 for everything else.
More sophisticated methods: use a lasso or a Bayesian prior to shrink effect sizes genome-wide.

And test how well that correlates with the phenotype in an external replication dataset. Square it and you get the “variance explained by the PRS”:

$$cor(y_{replication}, \hat{\beta}_{snp}^{T} G_{snp})^2 = h_{prs}^2$$

Estimating the SNP heritability using a polygenic risk score

Estimate a polygenic risk score, by trying to estimate the effect sizes (beta^hat):

$$\hat{\beta}_{snp}^{T} G_{snp}$$

Crude method: estimate betas for genome-wide significant hits and set beta = 0 for everything else.
More sophisticated methods: use a lasso or a Bayesian prior to shrink effect sizes genome-wide.

And test how well that correlates with the phenotype in an external replication dataset. Square it and you get the "variance explained by the PRS":

$$cor(y_{replication}, \hat{\beta}_{snp}^{T} G_{snp})^2 = h_{prs}^2 < h_{snp}^2$$

But this will be less than the true SNP heritability, as inaccuracy in beta^hat introduces error and reduces the correlation.

LD Score Regression – a better way of estimating SNP heritability

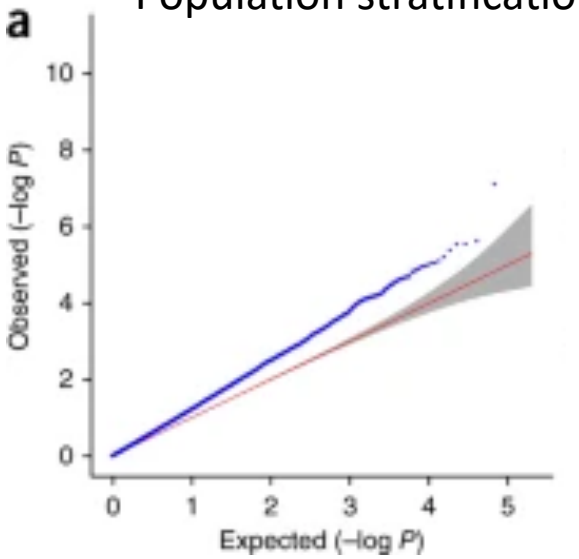
- We would like to measure the total amount of signal in the GWAS (eg by average p-value or chi-square statistic).
  - But this can be driven by real signal or population stratification

Bulik-Sullivan et al (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295

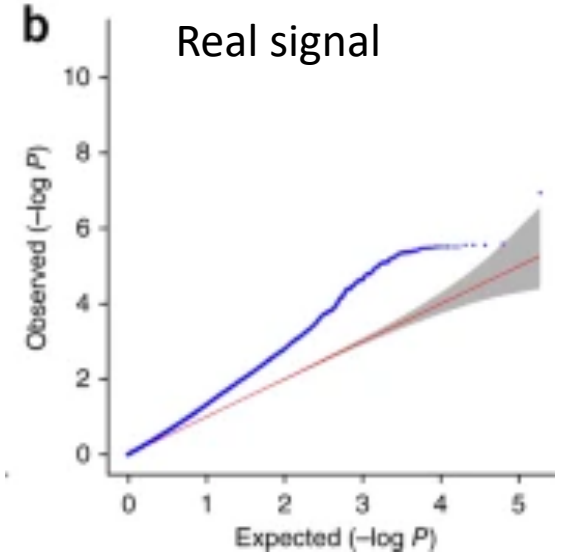
LD Score Regression – a better way of estimating SNP heritability

- We would like to measure the total amount of signal in the GWAS (eg by average p-value or chi-square statistic).
  - But this can be driven by real signal or population stratification

Population stratification



Real signal



Bulik-Sullivan et al (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Variant	LD Score
rs1	
rs2	
rs3	
rs4	

LD scores measure how much potential each variant has for tagging causal variants. The higher the LD score, the more true signal we expect it to tag, and the larger we expect it's test statistic (or the smaller we expect its p-value) to be.

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Variant	LD Score
rs1	1^2 + 0.1^2 + 0.05^2 + 0^2 = 1.0125
rs2	
rs3	
rs4	

LD scores measure how much potential each variant has for tagging causal variants. The higher the LD score, the more true signal we expect it to tag, and the larger we expect it's test statistic (or the smaller we expect its p-value) to be.

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Variant	LD Score
rs1	1^2 + 0.1^2 + 0.05^2 + 0^2 = 1.0125
rs2	0.1^2 + 1^2 + 0.9^2 + 0.85^2 = 2.5425
rs3	0.05^2+0.9^2+1^2+0.9^2 = 2.6225
rs4	0^2+0.85^2+0.9^2 + 1^2 = 2.5325

LD scores measure how much potential each variant has for tagging causal variants. The higher the LD score, the more true signal we expect it to tag, and the larger we expect it's test statistic (or the smaller we expect its p-value) to be.

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Variant	LD Score
rs1	1^2 + 0.1^2 + 0.05^2 + 0^2 = 1.0125
rs2	0.1^2 + 1^2 + 0.9^2 + 0.85^2 = 2.5425
rs3	0.05^2+0.9^2+1^2+0.9^2 = 2.6225
rs4	0^2+0.85^2+0.9^2 + 1^2 = 2.5325

Small LD score, only tags itself and no other causal variants

Large LD score, tags itself and two other causal variants.

LD scores measure how much potential each variant has for tagging causal variants. The higher the LD score, the more true signal we expect it to tag, and the larger we expect it's test statistic (or the smaller we expect its p-value) to be.

LD Scores

LD matrix:

	rs1	rs2	rs3	rs4
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

Variant	LD Score
rs1	1^2 + 0.1^2 + 0.05^2 + 0^2 = 1.0125
rs2	0.1^2 + 1^2 + 0.9^2 + 0.85^2 = 2.5425
rs3	0.05^2+0.9^2+1^2+0.9^2 = 2.6225
rs4	0^2+0.85^2+0.9^2 + 1^2 = 2.5325

Small LD score, only tags itself and no other causal variants

Large LD score, tags itself and two other causal variants.

LD scores measure how much potential each variant has for tagging causal variants. The higher the LD score, the more true signal we expect it to tag, and the larger we expect it's test statistic (or the smaller we expect its p-value) to be.

But more than that – the higher the heritability, the larger the slope between LD score and test statistic:

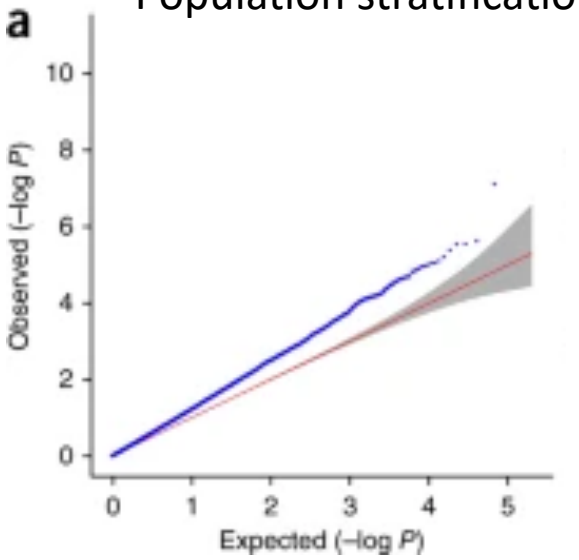
$$E[\chi^2|\ell_i] = h^2_{snp}\ell_i\frac{N}{M} + Na + 1$$

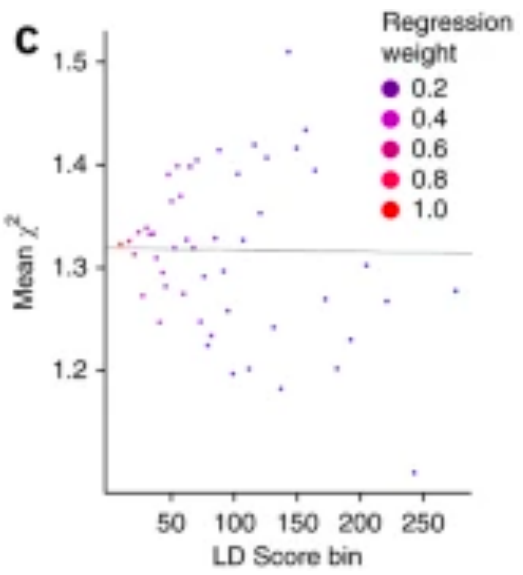
χ2 = test statistic, N = sample size, M = number of SNPs, a = confounding

LD Score Regression – a better way of estimating SNP heritability

- We would like to measure the total amount of signal in the GWAS (eg by average p-value or chi-square statistic).
  - But this can be driven by real signal or population stratification

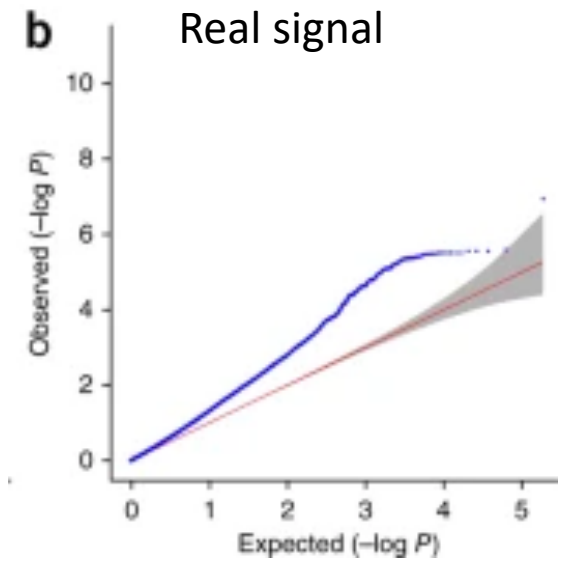
Population stratification

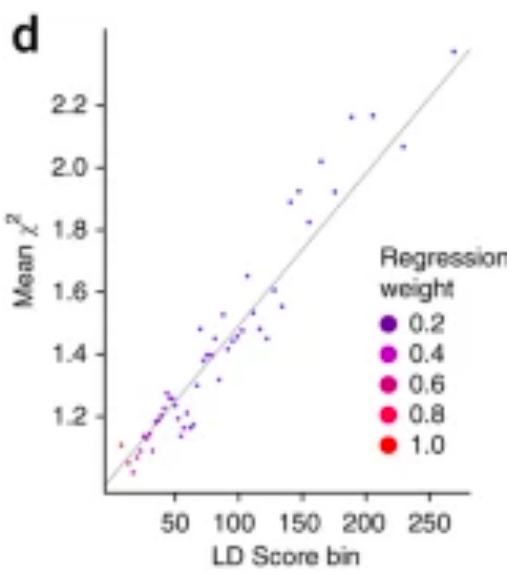




h2_snp = 0%

Real signal

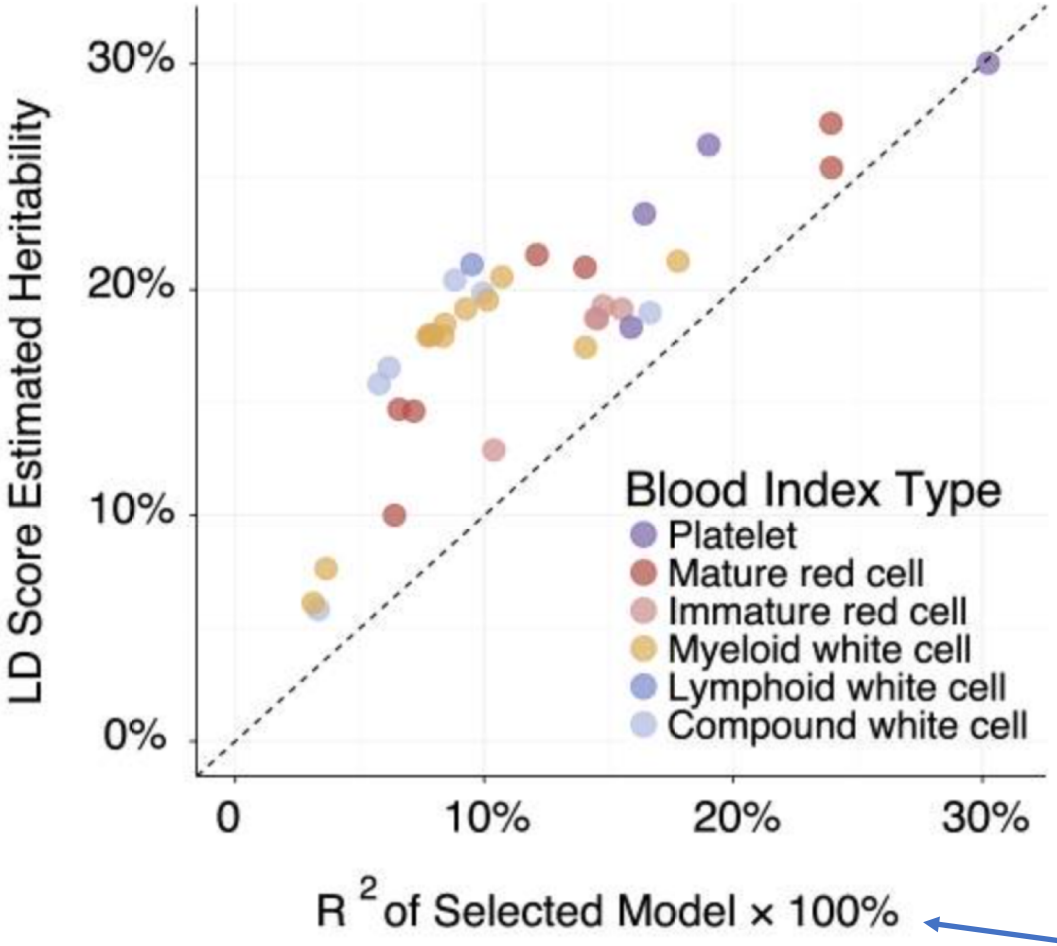




h2_snp = 33%

Back to the paper

This is h^2_snp



This is h^2_prs

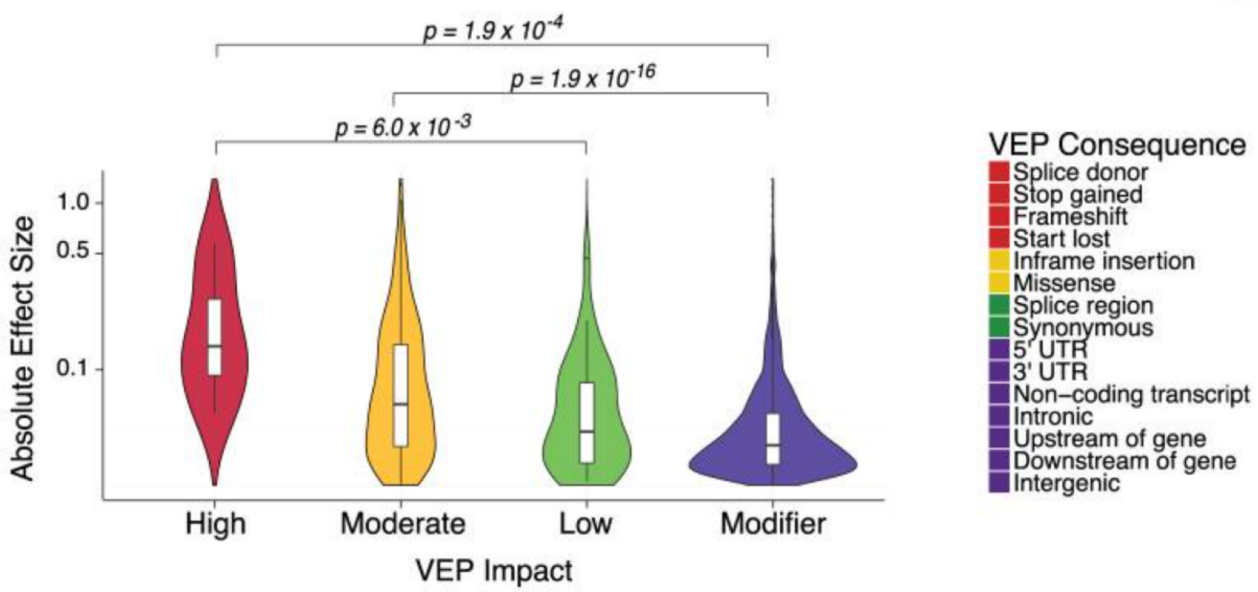
SNP heritability ranges from 5-30%. Polygenic risk scores (based on significant associations) capture a lot, but far from all, of this.

Partitioned heritability

- Often we are interested in how much heritability is explained by different type of variants:

Partitioned heritability

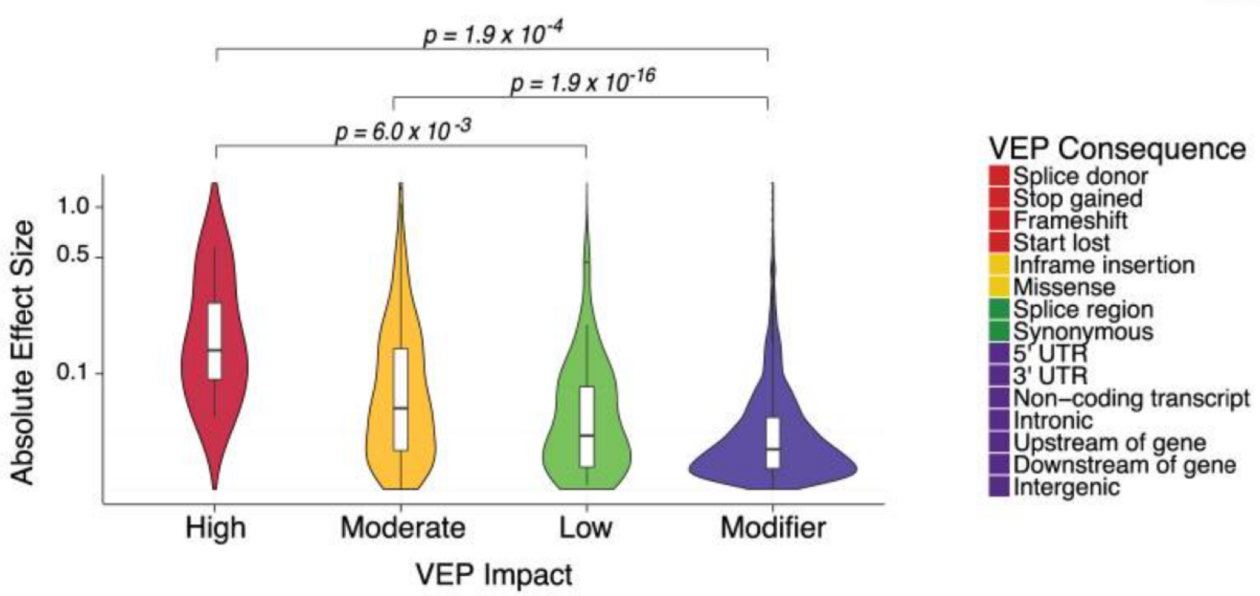
- Often we are interested in how much heritability is explained by different type of variants:



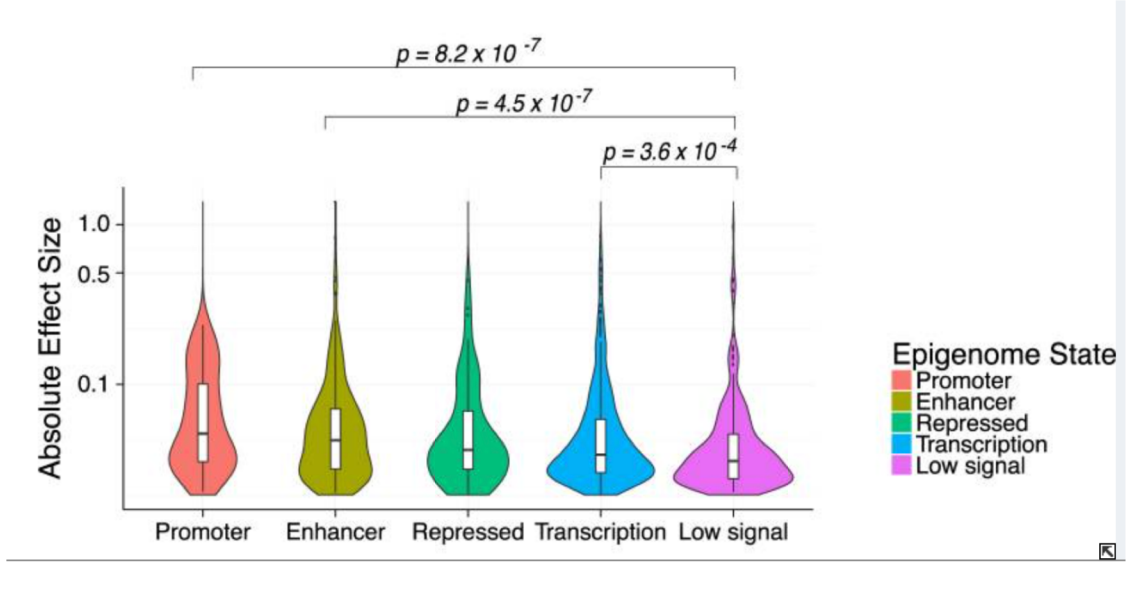
Eg different consequences of the mutation on nearby genes

Partitioned heritability

- Often we are interested in how much heritability is explained by different type of variants:



Eg different consequences of the mutation on nearby genes



Or variants that lie in different types of gene regulatory region.

Partitioned heritability

$$y = \beta^T_{coding} G_{coding} + \beta^T_{noncoding} G_{noncoding} + e$$

Genetic contribution of coding variants

Genetic contribution of coding variants

$$y \sim N(0,1)$$
$$\beta^T_{coding} G_{coding} \sim N(0, h^2_{coding})$$
$$\beta^T_{noncoding} G_{noncoding} \sim N(0, h^2_{noncoding})$$
$$e \sim N(0, 1 - h^2_{coding} - h^2_{noncoding})$$

Partitioned heritability

$$y = \beta^T_{coding} G_{coding} + \beta^T_{noncoding} G_{noncoding} + e$$

Genetic contribution of coding variants

Genetic contribution of coding variants

$$y \sim N(0, 1)$$
$$\beta^T_{coding} G_{coding} \sim N(0, h^2_{coding})$$
$$\beta^T_{noncoding} G_{noncoding} \sim N(0, h^2_{noncoding})$$
$$e \sim N(0, 1 - h^2_{coding} - h^2_{noncoding})$$

We break the heritability down into contributions from each category. We want to estimate the heritability of each category (h2_coding, h2_noncoding, etc).

Partitioned LD Scores

LD matrix:

	rs1 (coding)	rs2 (noncoding)	rs3 (noncoding)	rs4 (noncoding)
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

Partitioned LD Scores

LD matrix:

	rs1 (coding)	rs2 (noncoding)	rs3 (noncoding)	rs4 (noncoding)
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell(j,C) = \sum_{k \in C} r_{jk}^2$$

Partitioned LD score for category C

Variant	Coding LD score	Noncoding LD score
rs1		
rs2		
rs3		
rs4		

Partitioned LD scores measure how much each variant tags different classes of variant.

Partitioned LD Scores

LD matrix:

	rs1 (coding)	rs2 (noncoding)	rs3 (noncoding)	rs4 (noncoding)
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell(j,C) = \sum_{k \in C} r_{jk}^2$$

Partitioned LD score for category C

Variant	Coding LD score	Noncoding LD score
rs1	1^2 = 1	0.1^2 + 0.05^2 + 0^2 = 0.0125
rs2		
rs3		
rs4		

Partitioned LD scores measure how much each variant tags different classes of variant.

Partitioned LD Scores

LD matrix:

	rs1 (coding)	rs2 (noncoding)	rs3 (noncoding)	rs4 (noncoding)
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores:

$$\ell(j,C) = \sum_{k \in C} r_{jk}^2$$

Partitioned LD score for category C

Variant	Coding LD score	Noncoding LD score
rs1	1^2 = 1	0.1^2 + 0.05^2 + 0^2 = 0.0125
rs2	0.1^2 = 0.01	1^2 + 0.9^2 + 0.85^2 = 2.5325
rs3	0.05^2 = 0.00025	0.9^2 + 1^2 + 0.9^2 = 2.62
rs4	0^2 = 0	0.85^2 + 0.9^2 + 1^2 = 2.5325

Partitioned LD scores measure how much each variant tags different classes of variant.

Partitioned LD Scores

LD matrix:

	rs1 (coding)	rs2 (noncoding)	rs3 (noncoding)	rs4 (noncoding)
rs1	1	0.1	0.05	0
rs2	0.1	1	0.9	0.85
rs3	0.05	0.9	1	0.9
rs4	0	0.85	0.9	1

LD scores: $\ell(j,C) = \sum_{k \in C} r_{jk}^2$ Partitioned LD score for category C

Variant	Coding LD score	Noncoding LD score
rs1	1^2 = 1	0.1^2 + 0.05^2 + 0^2 = 0.0125
rs2	0.1^2 = 0.01	1^2 + 0.9^2 + 0.85^2 = 2.5325
rs3	0.05^2 = 0.00025	0.9^2 + 1^2 + 0.9^2 = 2.62
rs4	0^2 = 0	0.85^2 + 0.9^2 + 1^2 = 2.5325

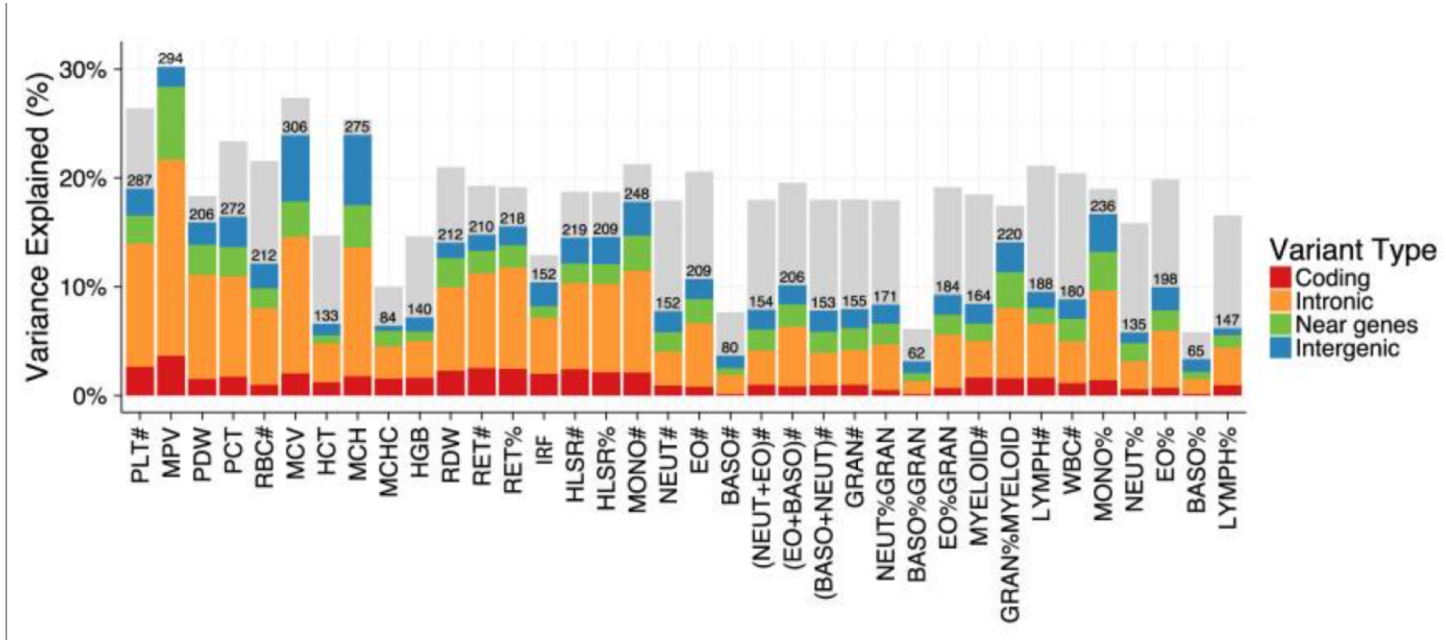
Partitioned LD scores measure how much each variant tags different classes of variant.

We can now estimate the two different slopes:

$$E[\chi_i^2] = h_{coding}^2 \ell(i, coding)\frac{N}{M} + h_{noncoding}^2 \ell(i, noncoding)\frac{N}{M} + Na + 1$$

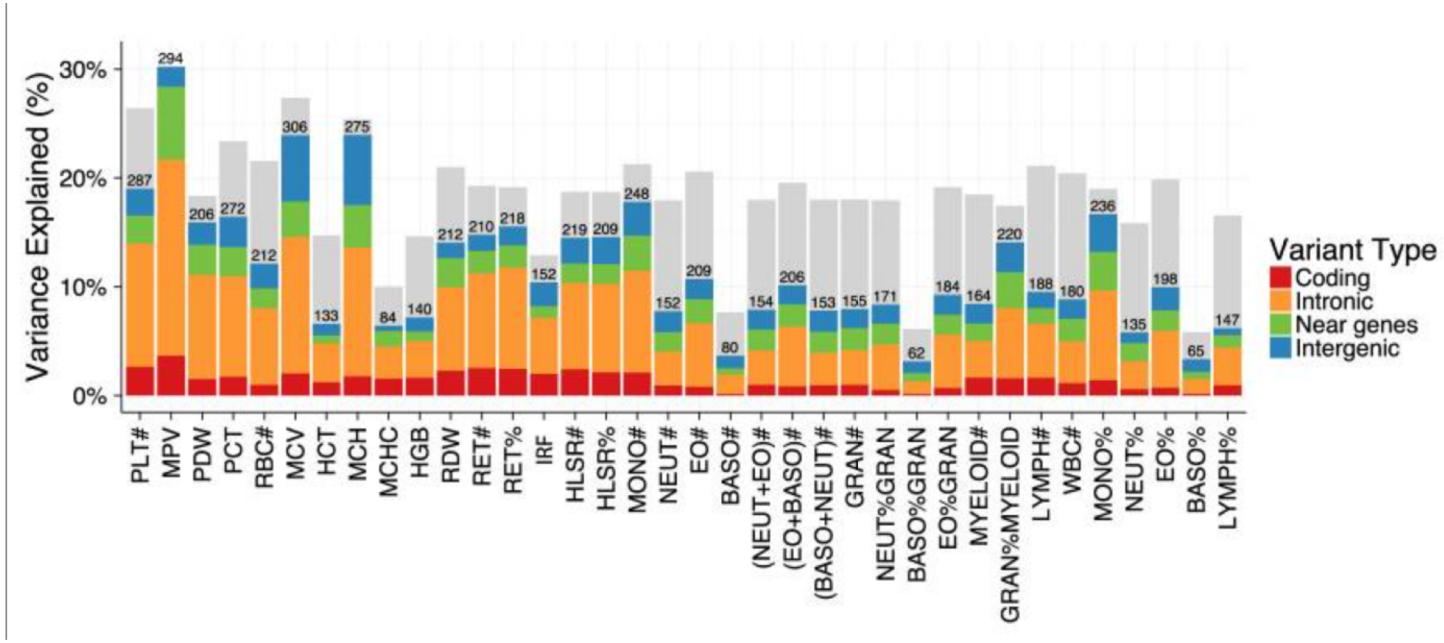
χ2 = test statistic, N = sample size, M = number of SNPs, a = confounding

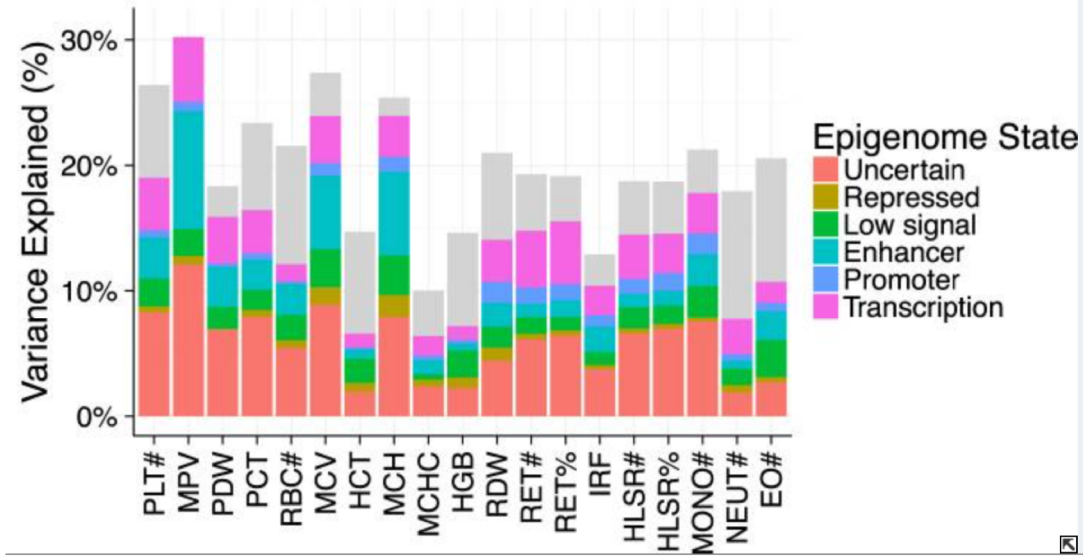
Back to the paper



Most of the heritability in blood traits is driving by intronic variation

Back to the paper





Most of the heritability in blood traits is driving by intronic variation

and by variants in enhancers and transcribed regions.

Running these yourself

- We do not have a practical on LD Score regression, but the LDSC authors have a number of good tutorials with real, publicly available data on their wiki:

https://github.com/bulik/ldsc/wiki

Fine-mapping

Swapping out for a new paper

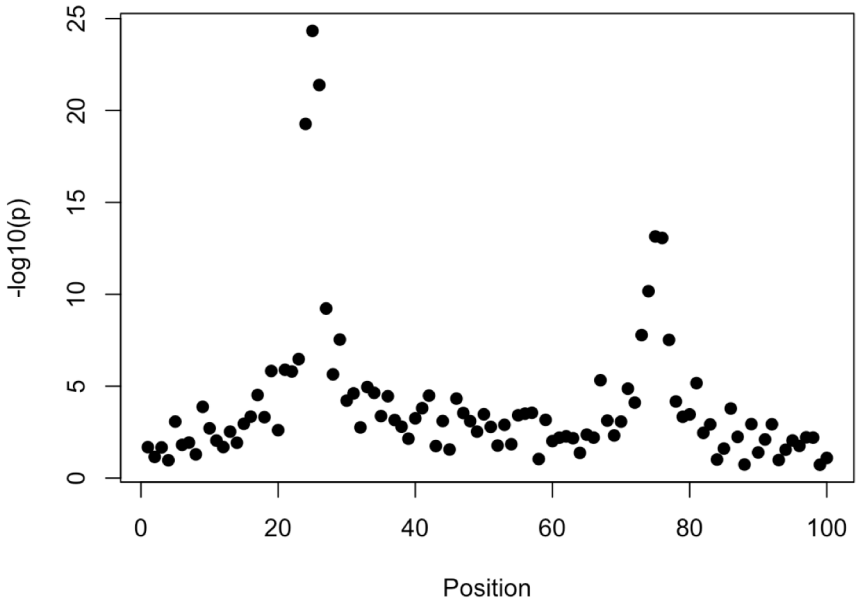
ARTICLES
https://doi.org/10.1038/s41588-021-00880-5
nature genetics

Check for updates

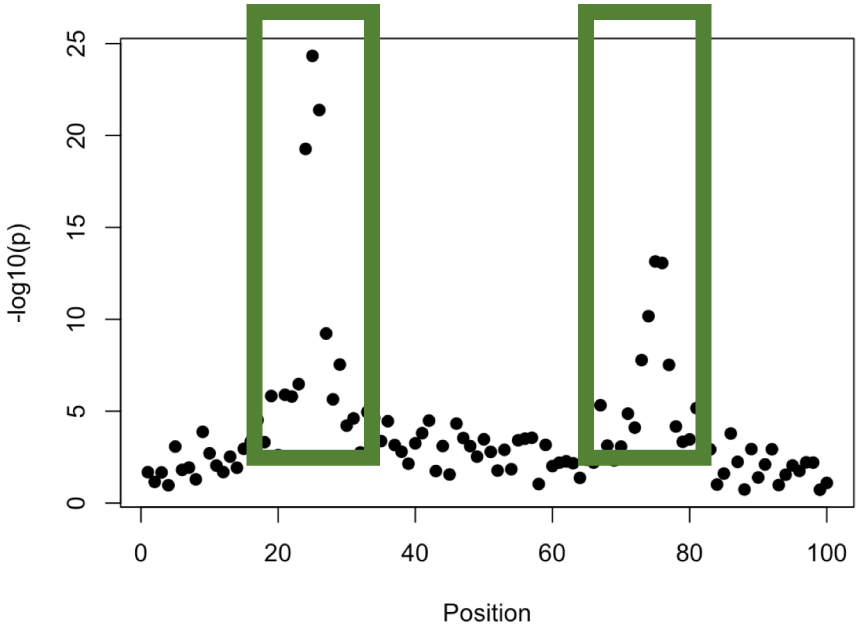
Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes

Catherine C. Robertson[1,2,29], Jamie R. J. Inshaw[3,29], Suna Onengut-Gumuscu[1,4], Wei-Min Chen[1,4], David Flores Santa Cruz[3], Hanzhi Yang[1], Antony J. Cutler[3], Daniel J. M. Crouch[3], Emily Farber[1], S. Louis Bridges Jr[5,6], Jeffrey C. Edberg[7], Robert P. Kimberly[7], Jane H. Buckner[8], Panos Deloukas[9,10],

Aims of fine-mapping

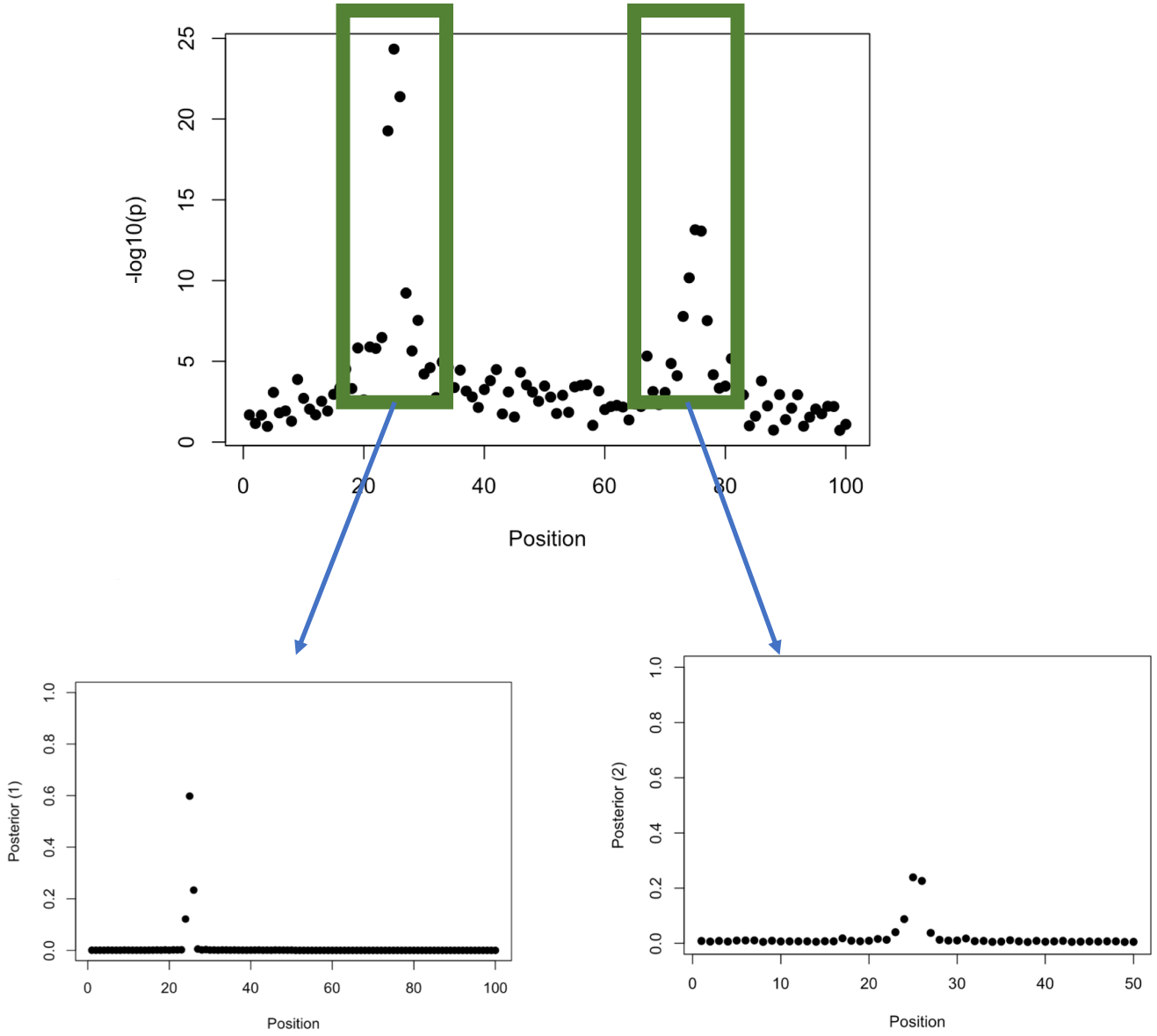


Aims of fine-mapping



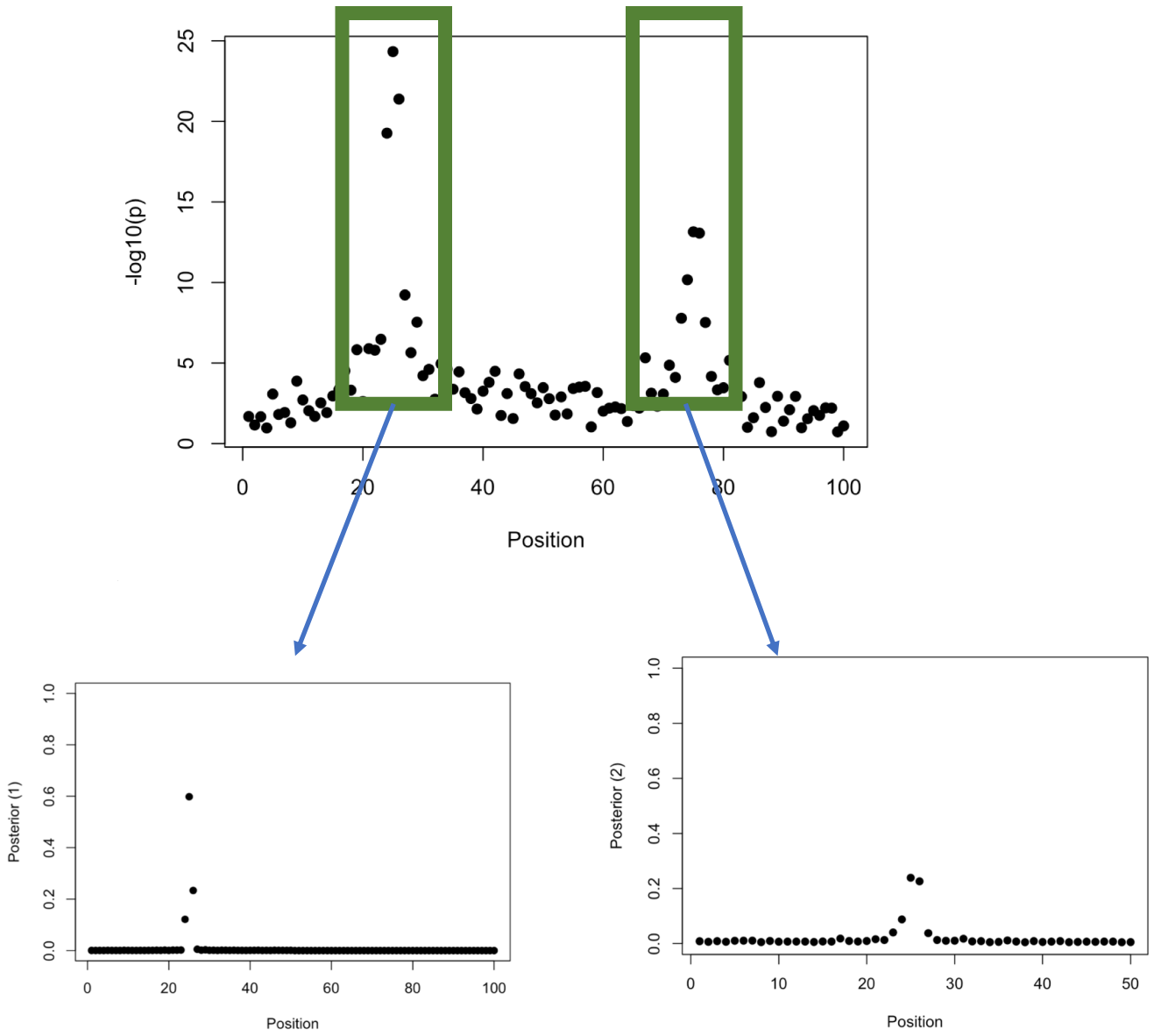
1. Identify the number of independent signals (i.e. the number of causal variants) in the region.

Aims of fine-mapping



1. Identify the number of independent signals (i.e. the number of causal variants) in the region.
2. Identify the candidates for the actual causal variant for each signal.

Aims of fine-mapping



1. Identify the number of independent signals (i.e. the number of causal variants) in the region.
2. Identify the candidates for the actual causal variant for each signal.
3. Identify possible functions of these causal variants

Maller et al fine-mapping

Maller et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 44(12): 1294–1301

Maller et al fine-mapping

In Maller-style mapping, we assume that there is only one causal variant, and thus we can consider each variant one-at-a-time:

$$P(C = i|D) = \frac{P(D|C = i)P(C = i)}{\sum_j P(D|C = j)P(C = j)}$$

Maller et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 44(12): 1294–1301

Maller et al fine-mapping

In Maller-style mapping, we assume that there is only one causal variant, and thus we can consider each variant one-at-a-time:

$$P(C = i|D) = \frac{P(D|C = i)P(C = i)}{\sum_j P(D|C = j)P(C = j)}$$

$$= \frac{BF_i}{\sum_j BF_j}$$

Assume P(C = i) is a constant for all i

Maller et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 44(12): 1294–1301

Maller et al fine-mapping

In Maller-style mapping, we assume that there is only one causal variant, and thus we can consider each variant one-at-a-time:

$$P(C = i|D) = \frac{P(D|C = i)P(C = i)}{\sum_j P(D|C = j)P(C = j)}$$

$$= \frac{BF_i}{\sum_j BF_j}$$

Assume P(C = i) is a constant for all i

Where BF is the Bayes factor:

$$BF_i = \frac{f(\hat{\beta}_i|\mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i|\mu = 0, \sigma^2 = se_i^2)}$$

Maller et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 44(12): 1294–1301

Maller et al fine-mapping

In Maller-style mapping, we assume that there is only one causal variant, and thus we can consider each variant one-at-a-time:

$$P(C = i|D) = \frac{P(D|C = i)P(C = i)}{\sum_j P(D|C = j)P(C = j)}$$

$$= \frac{BF_i}{\sum_j BF_j}$$

Assume P(C = i) is a constant for all i

Where BF is the Bayes factor:

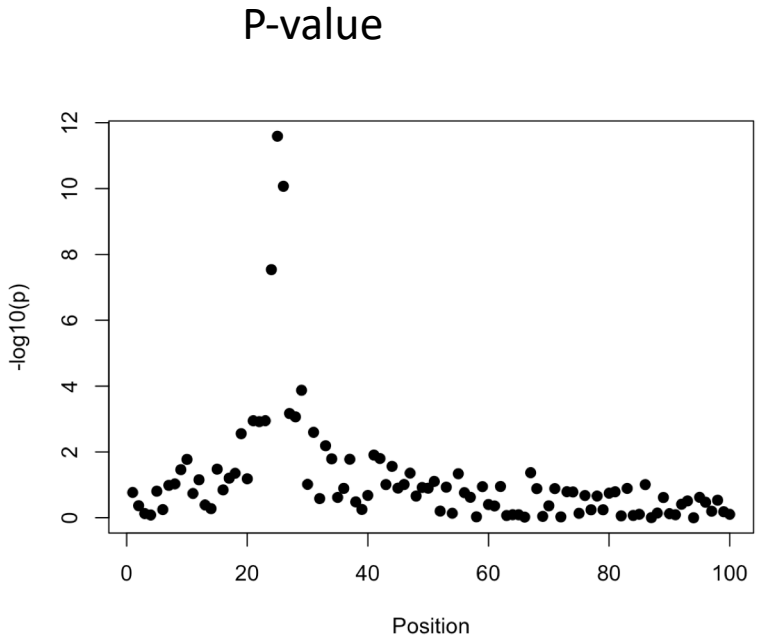
$$BF_i = \frac{f(\hat{\beta}_i|\mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i|\mu = 0, \sigma^2 = se_i^2)}$$

The likelihood of seeing the observed beta_i given a non-zero effect size prior sigma_0

The likelihood of seeing the observed beta_i under the null

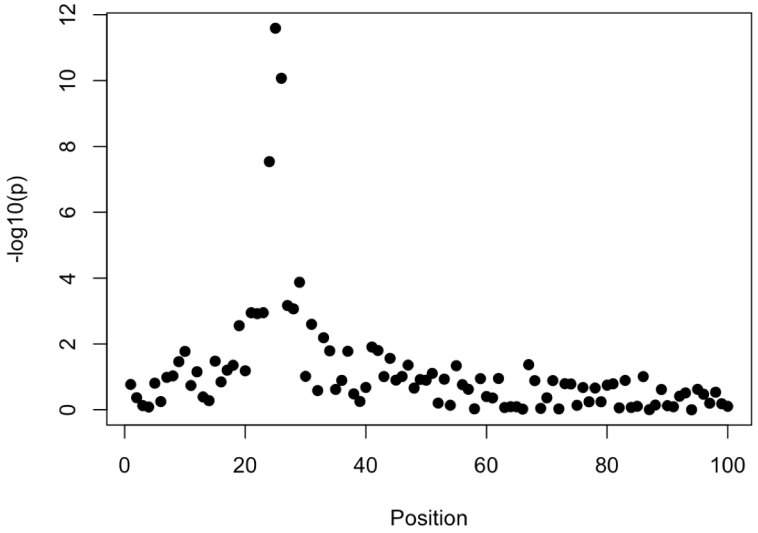
Maller et al (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 44(12): 1294–1301

Maller et al fine-mapping

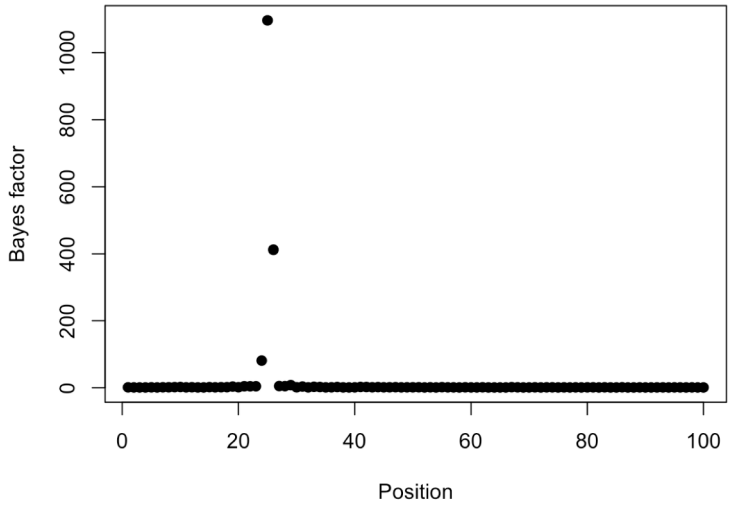


Maller et al fine-mapping

P-value

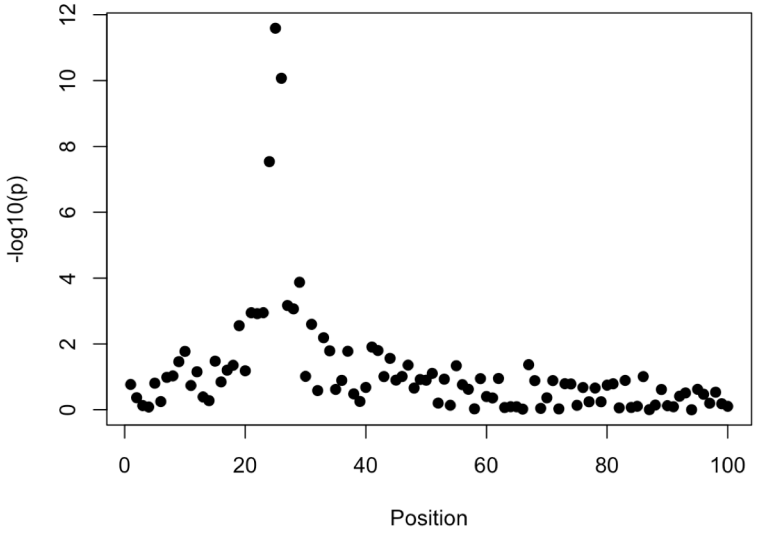


Bayes factor

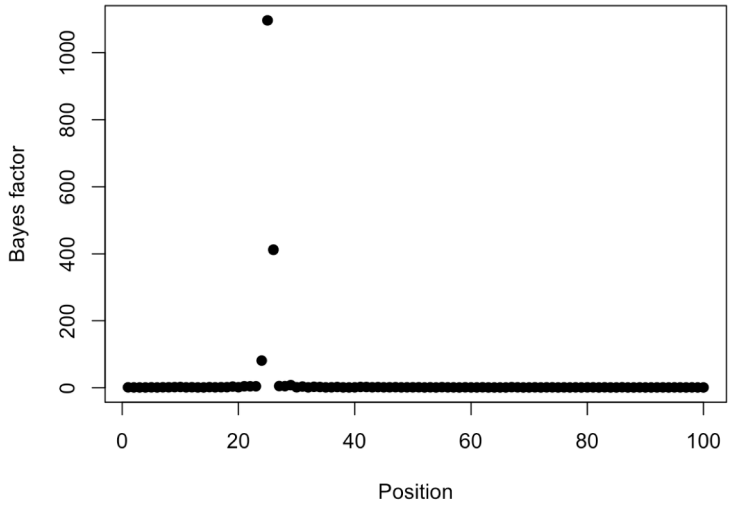


Maller et al fine-mapping

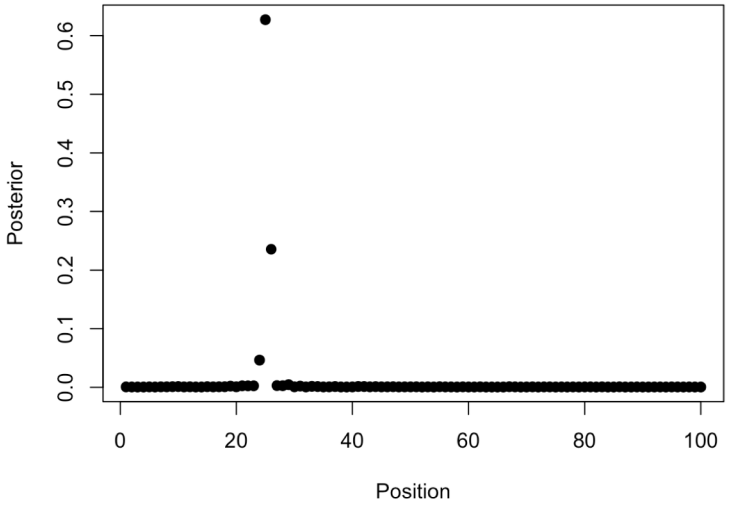
P-value



Bayes factor



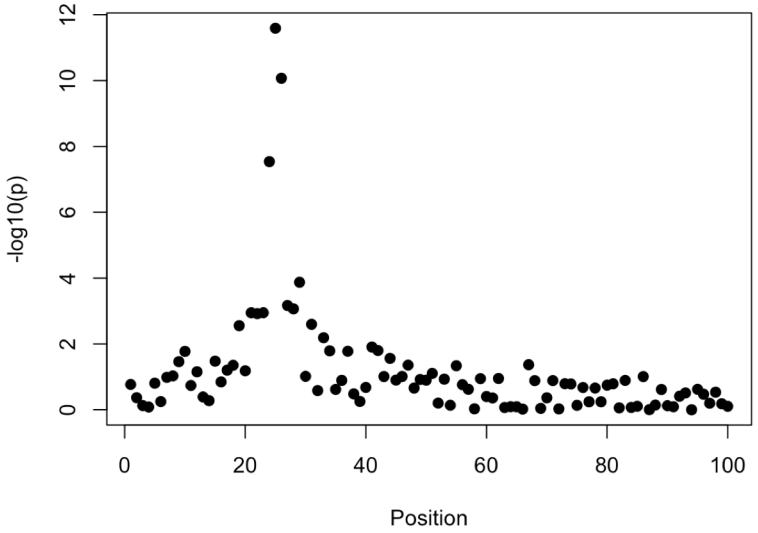
Posterior:



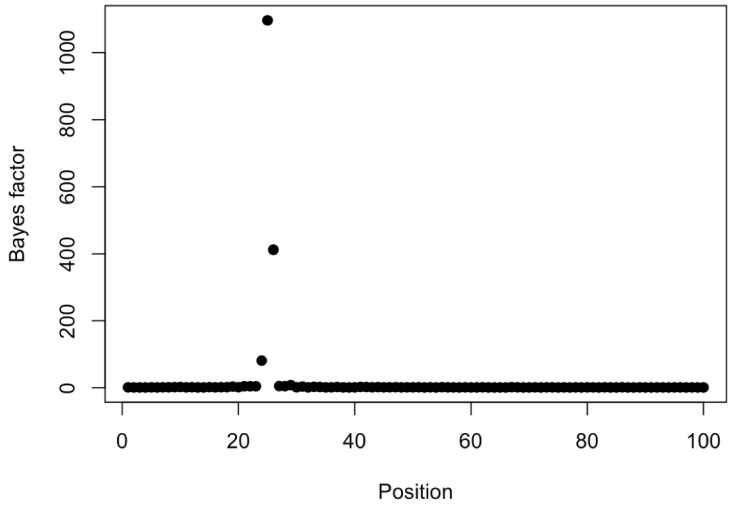
A note on **credible sets**: Rank the SNPs by posterior, and go down the list adding them up. When you go over 95%, that is your 95% credible set. There is at least a 95% chance that the true causal variant is in this set.

Maller et al fine-mapping

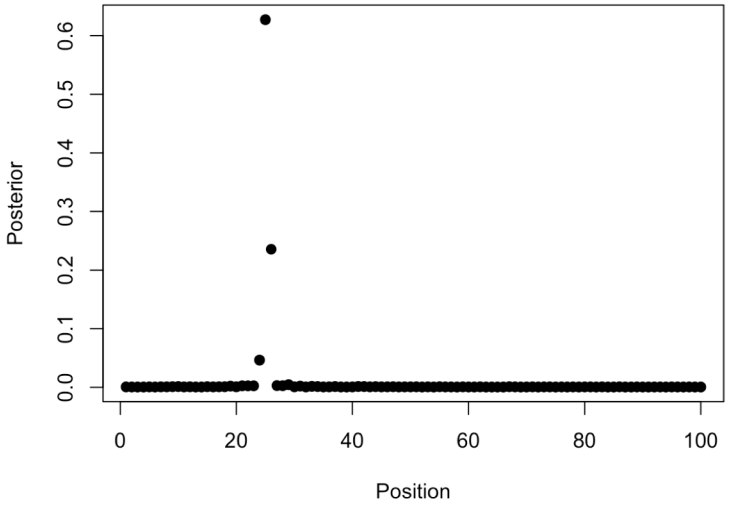
P-value



Bayes factor



Posterior:

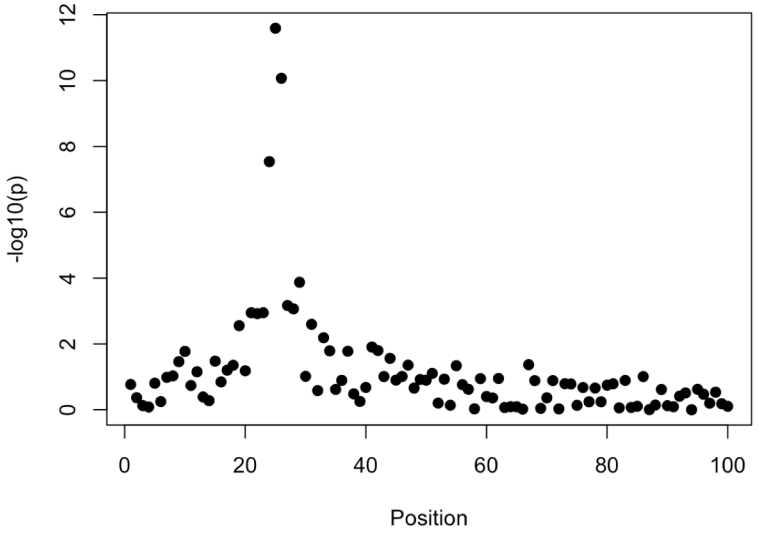


A note on **credible sets**: Rank the SNPs by posterior, and go down the list adding them up. When you go over 95%, that is your 95% credible set. There is at least a 95% chance that the true causal variant is in this set.

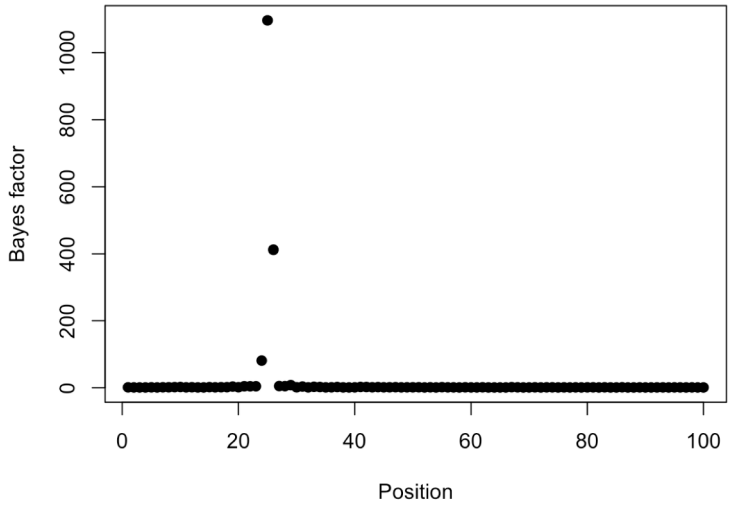
Variant	Posterior	
rs1	0.06	
rs2	0.65	
rs3	0.25	
rs4	0.02	

Maller et al fine-mapping

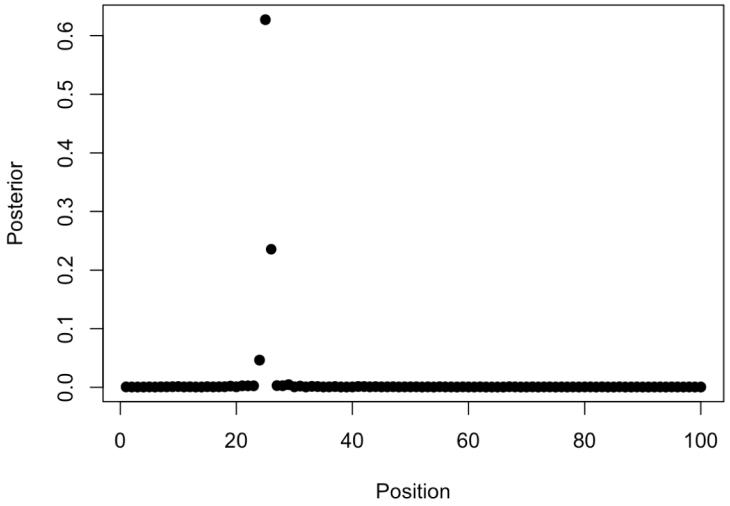
P-value



Bayes factor



Posterior:

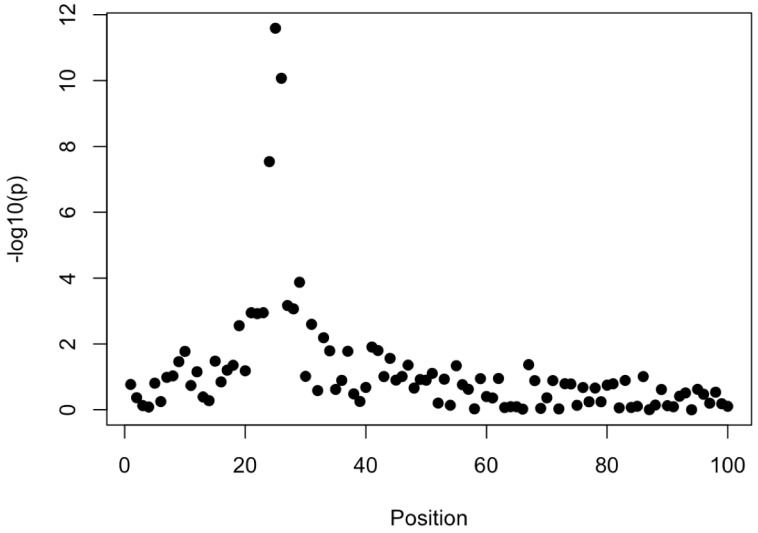


A note on **credible sets**: Rank the SNPs by posterior, and go down the list adding them up. When you go over 95%, that is your 95% credible set. There is at least a 95% chance that the true causal variant is in this set.

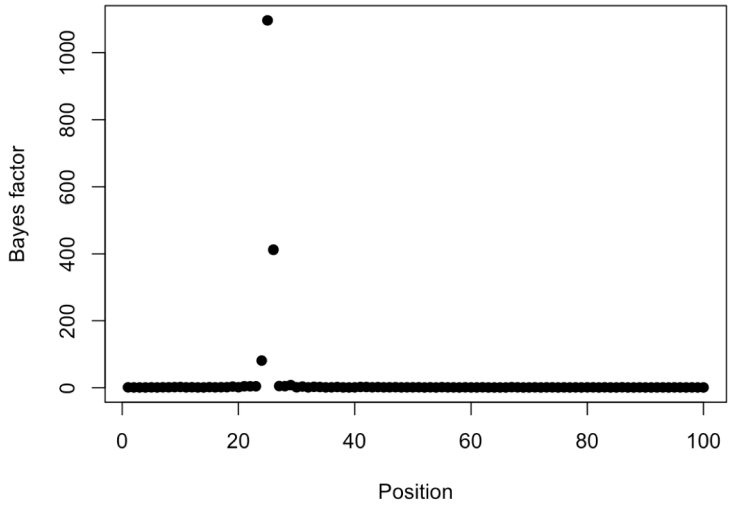
Variant	Posterior	
rs2	0.65	
rs3	0.25	
rs1	0.06	
rs4	0.02	

Maller et al fine-mapping

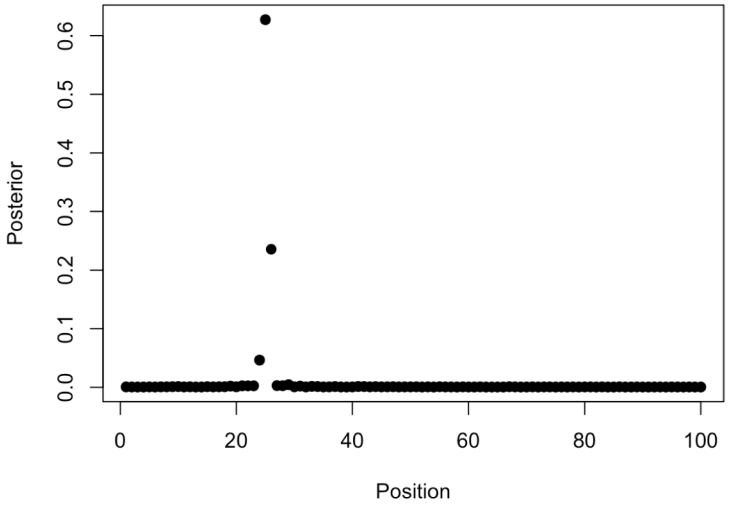
P-value



Bayes factor



Posterior:

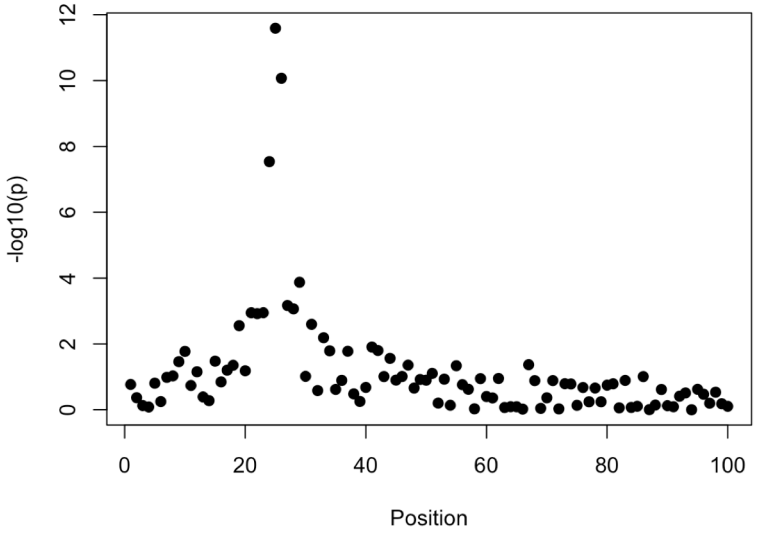


A note on **credible sets**: Rank the SNPs by posterior, and go down the list adding them up. When you go over 95%, that is your 95% credible set. There is at least a 95% chance that the true causal variant is in this set.

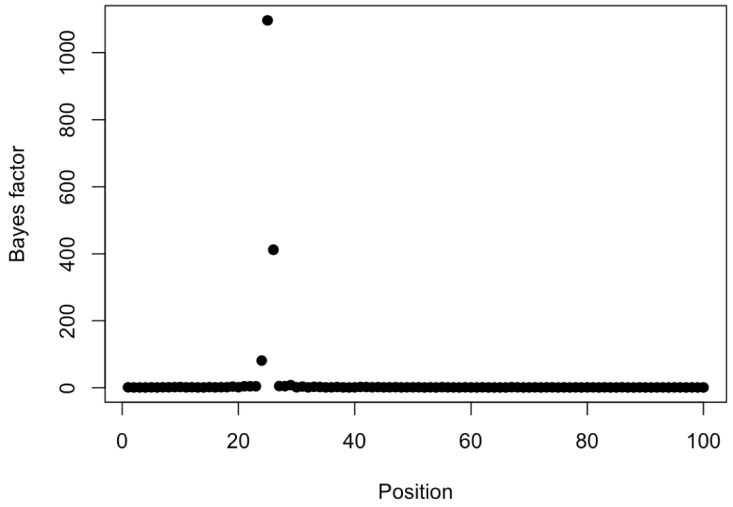
Variant	Posterior	Cumsum
rs2	0.65	0.65
rs3	0.25	0.90
rs1	0.06	0.96
rs4	0.02	0.98

Maller et al fine-mapping

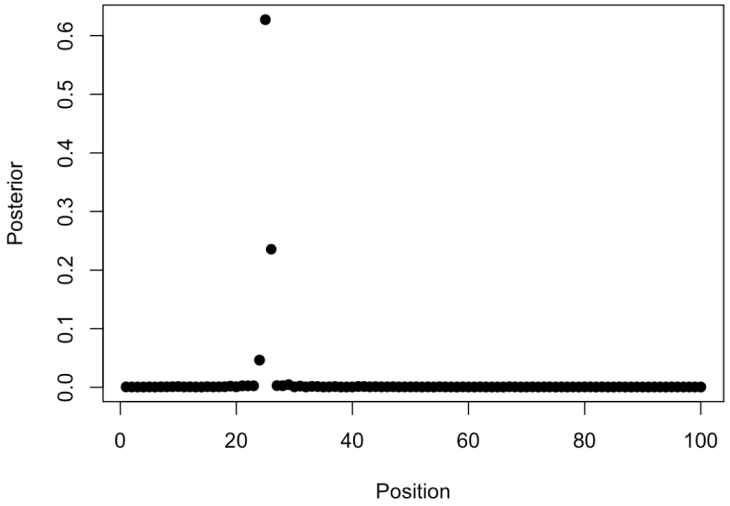
P-value



Bayes factor



Posterior:



A note on **credible sets**: Rank the SNPs by posterior, and go down the list adding them up. When you go over 95%, that is your 95% credible set. There is at least a 95% chance that the true causal variant is in this set.

Credible set

Variant	Posterior	Cumsum
rs2	**0.65**	**0.65**
rs3	**0.25**	**0.90**
rs1	**0.06**	**0.96**
rs4	0.02	0.98

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)				
f(beta_i | mu = 0, sigma = se^2)				
BF_i				
Posterior_i				

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)	dnorm(0.4,0,sqrt(0.05^2 + 0.2^2)) = 0.295	dnorm(0.41,0,sqrt(0.05^2 + 0.2^2)) = 0.268	dnorm(0.3,0,sqrt(0.1^2 + 0.2^2)) = 0.725	dnorm(0.1,0,sqrt(0.1^2 + 0.2^2)) = 1.614
f(beta_i | mu = 0, sigma = se^2)				
BF_i				
Posterior_i				

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)	dnorm(0.4,0,sqrt(0.05^2 + 0.2^2)) = 0.295	dnorm(0.41,0,sqrt(0.05^2 + 0.2^2)) = 0.268	dnorm(0.3,0,sqrt(0.1^2 + 0.2^2)) = 0.725	dnorm(0.1,0,sqrt(0.1^2 + 0.2^2)) = 1.614
f(beta_i | mu = 0, sigma = se^2)	dnorm(0.4,0,sqrt(0.05^2)) = 1.01e-13	dnorm(0.42,0,sqrt(0.05^2)) = 3.80e-15	dnorm(0.3,0,sqrt(0.1^2)) = 0.0443	dnorm(0.1,0,sqrt(0.1^2)) = 2.420
BF_i				
Posterior_i				

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)	dnorm(0.4,0,sqrt(0.05^2 + 0.2^2)) = 0.295	dnorm(0.41,0,sqrt(0.05^2 + 0.2^2)) = 0.268	dnorm(0.3,0,sqrt(0.1^2 + 0.2^2)) = 0.725	dnorm(0.1,0,sqrt(0.1^2 + 0.2^2)) = 1.614
f(beta_i | mu = 0, sigma = se^2)	dnorm(0.4,0,sqrt(0.05^2)) = 1.01e-13	dnorm(0.42,0,sqrt(0.05^2)) = 3.80e-15	dnorm(0.3,0,sqrt(0.1^2)) = 0.0443	dnorm(0.1,0,sqrt(0.1^2)) = 2.420
BF_i	0.295/1.01e-13 = 2.92e12	0.243/3.80e-15 = 6.39e13	0.725/0.0443 = 16.37	1.614/2.420 = 0.667
Posterior_i				

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)	dnorm(0.4,0,sqrt(0.05^2 + 0.2^2)) = 0.295	dnorm(0.41,0,sqrt(0.05^2 + 0.2^2)) = 0.268	dnorm(0.3,0,sqrt(0.1^2 + 0.2^2)) = 0.725	dnorm(0.1,0,sqrt(0.1^2 + 0.2^2)) = 1.614
f(beta_i | mu = 0, sigma = se^2)	dnorm(0.4,0,sqrt(0.05^2)) = 1.01e-13	dnorm(0.42,0,sqrt(0.05^2)) = 3.80e-15	dnorm(0.3,0,sqrt(0.1^2)) = 0.0443	dnorm(0.1,0,sqrt(0.1^2)) = 2.420
BF_i	0.295/1.01e-13 = 2.92e12	0.243/3.80e-15 = 6.39e13	0.725/0.0443 = 16.37	1.614/2.420 = 0.667
Posterior_i				

Sum of BFs = 6.682e13

A simple worked example

$$BF_i = \frac{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2 + \sigma_0^2)}{f(\hat{\beta}_i | \mu = 0, \sigma^2 = se_i^2)}$$

$$posterior_i = \frac{BF_i}{\sum_j BF_j}$$

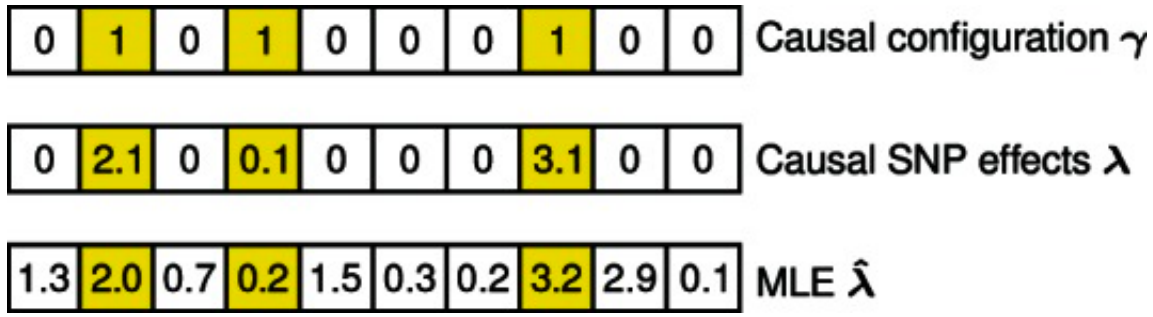
Variant:	rs1	rs2	rs3	rs4
Effect size (beta_i)	0.4	0.41	0.3	0.1
Standard error (se_i)	0.05	0.05	0.1	0.1
f(beta_i | mu = 0, sigma = se^2 + sigma_0^2)	dnorm(0.4,0,sqrt(0.05^2 + 0.2^2)) = 0.295	dnorm(0.41,0,sqrt(0.05^2 + 0.2^2)) = 0.268	dnorm(0.3,0,sqrt(0.1^2 + 0.2^2)) = 0.725	dnorm(0.1,0,sqrt(0.1^2 + 0.2^2)) = 1.614
f(beta_i | mu = 0, sigma = se^2)	dnorm(0.4,0,sqrt(0.05^2)) = 1.01e-13	dnorm(0.42,0,sqrt(0.05^2)) = 3.80e-15	dnorm(0.3,0,sqrt(0.1^2)) = 0.0443	dnorm(0.1,0,sqrt(0.1^2)) = 2.420
BF_i	0.295/1.01e-13 = 2.92e12	0.243/3.80e-15 = 6.39e13	0.725/0.0443 = 16.37	1.614/2.420 = 0.667
Posterior_i	2.92e12/6.39e13 = 0.0457	6.39e13/6.682e13 = 0.956	16.37/6.682e13 = 2.45e-13	0.667/6.682e13 = 9.98e-15

Sum of BFs = 6.682e13

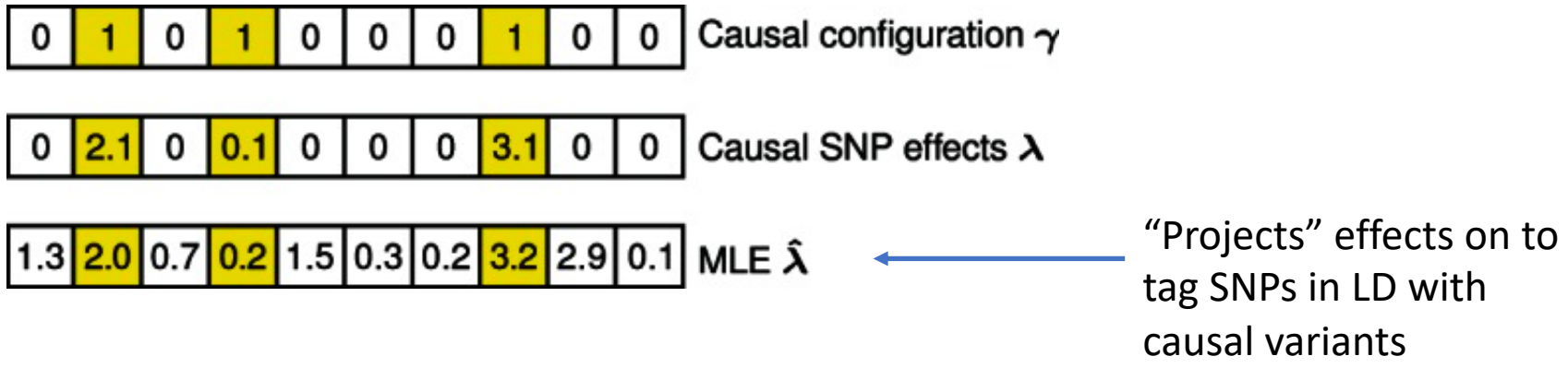
More complex fine-mapping

- There are often more than one causal variant at each locus, and we want to:
  - a) know how many there are
  - b) fine-map each causal variant while controlling for possible LD with other causal variants
- There are multiple techniques for fine-mapping in the presence of multiple causal variants
  - The paper uses GUESSFM, which is robust and well-behaved but requires access to complete genotype data
  - In the practical, we will use FINEMAP, which works just on summary statistics (betas, standard errors and an LD matrix).

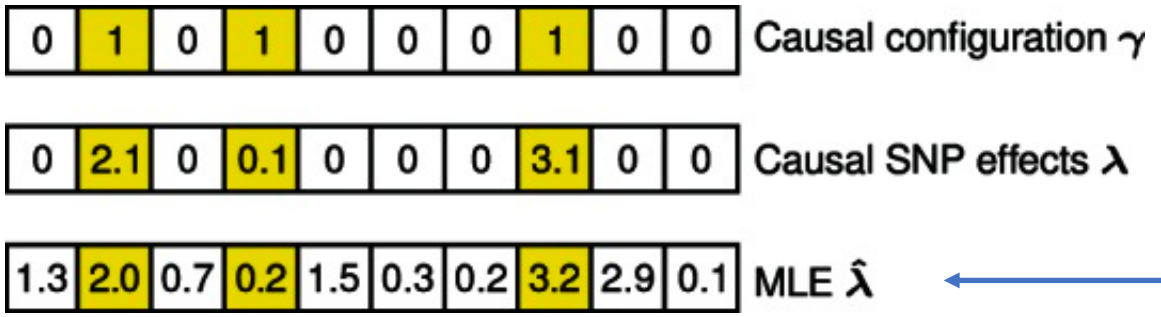
How FINEMAP works



How FINEMAP works



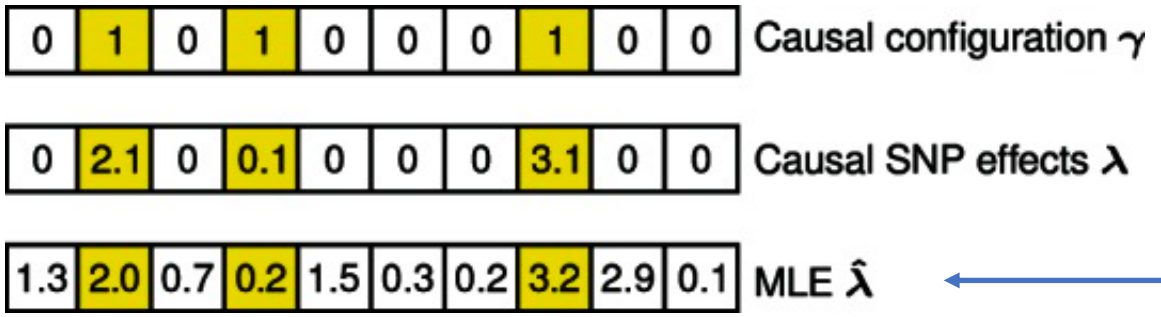
How FINEMAP works



"Projects" effects on to tag SNPs in LD with causal variants

The aim is to to calculate the posteriors for all plausible causal configurations.

How FINEMAP works



"Projects" effects on to tag SNPs in LD with causal variants

The aim is to to calculate the posteriors for all plausible causal configurations.

$$p(\boldsymbol{\gamma}) = p_k / \binom{m}{k} \quad \text{when} \sum_{\ell=1}^{m} \gamma_\ell = k.$$

Prior on the causal configuration

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathrm{N}(\boldsymbol{\lambda}|\mathbf{0}, s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma),$$

Prior on the effect sizes conditional on the causal configuration

How FINEMAP works

"Projects" effects on to tag SNPs in LD with causal variants

The aim is to to calculate the posteriors for all plausible causal configurations.

$$p(\boldsymbol{\gamma}) = p_k / \binom{m}{k} \quad \text{when} \sum_{\ell=1}^{m} \gamma_\ell = k.$$

Prior on the causal configuration

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathrm{N}(\boldsymbol{\lambda}|\mathbf{0}, s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma),$$

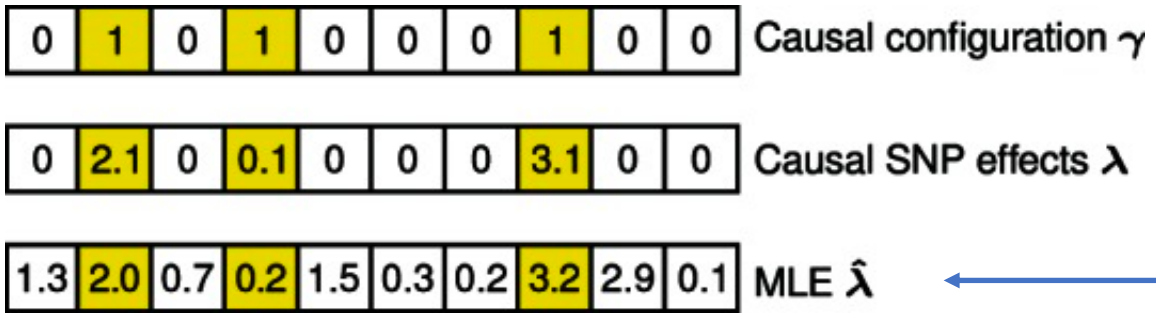
Prior on the effect sizes conditional on the causal configuration

$$p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}) = \int p(\mathbf{y}|\boldsymbol{\lambda}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\lambda}$$

$$= \mathcal{N}(\hat{\boldsymbol{\lambda}}|\mathbf{0}, \sigma^2 (n\mathbf{R})^{-1} + s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma)$$

The likelihood, marginalizing out the effect size

How FINEMAP works



"Projects" effects on to tag SNPs in LD with causal variants

The aim is to to calculate the posteriors for all plausible causal configurations.

$$p(\boldsymbol{\gamma}) = p_k / \binom{m}{k} \quad \text{when} \sum_{\ell=1}^{m} \gamma_\ell = k.$$

Prior on the causal configuration

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathrm{N}(\boldsymbol{\lambda}|\mathbf{0}, s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma),$$

Prior on the effect sizes conditional on the causal configuration

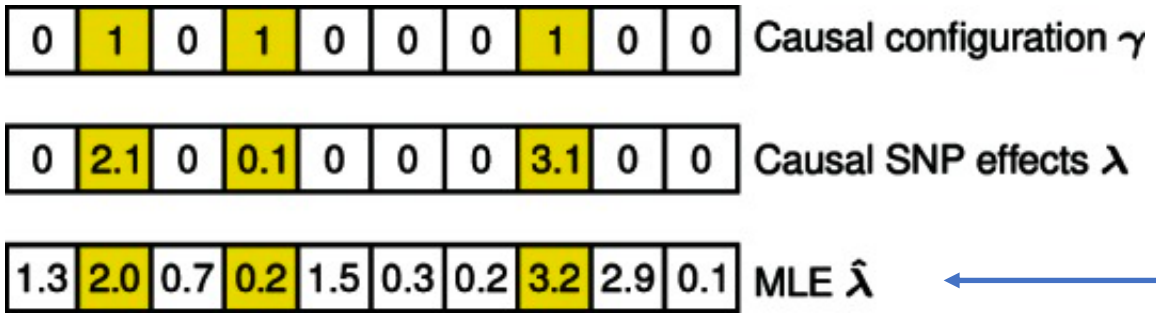
$$p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}) = \int p(\mathbf{y}|\boldsymbol{\lambda}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\lambda}$$
$$= \mathcal{N}(\hat{\boldsymbol{\lambda}}|\mathbf{0}, \sigma^2 (n\mathbf{R})^{-1} + s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma)$$

The likelihood, marginalizing out the effect size

$$p_1^*(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}) = \binom{m}{k}^{-1} p_k \times p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}),$$

The unnormalized posterior for a given causal configuration.

How FINEMAP works



"Projects" effects on to tag SNPs in LD with causal variants

The aim is to to calculate the posteriors for all plausible causal configurations.

$$p(\boldsymbol{\gamma}) = p_k / \binom{m}{k} \quad \text{when} \sum_{\ell=1}^{m} \gamma_\ell = k.$$

Prior on the causal configuration

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathrm{N}(\boldsymbol{\lambda}|\mathbf{0}, s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma),$$

Prior on the effect sizes conditional on the causal configuration

$$p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}) = \int p(\mathbf{y}|\boldsymbol{\lambda}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\lambda}$$
$$= \mathcal{N}(\hat{\boldsymbol{\lambda}}|\mathbf{0}, \sigma^2 (n\mathbf{R})^{-1} + s_\lambda^2 \sigma^2 \boldsymbol{\Delta}_\gamma)$$

The likelihood, marginalizing out the effect size

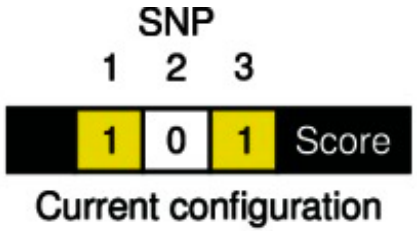
$$p_1^*(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}) = \binom{m}{k}^{-1} p_k \times p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}),$$

The unnormalized posterior for a given causal configuration.

Now we just need to calculate this for all possible causal configurations!
BUT For 10 SNPs -> 1024 configurations. 80 SNPs -> more configurations than there are stars in the universe.

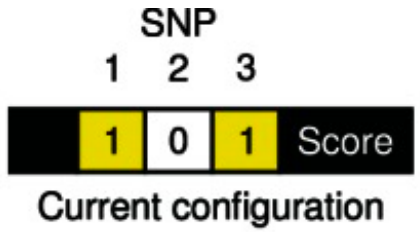
The shotgun stochastic search

Make some initial guess

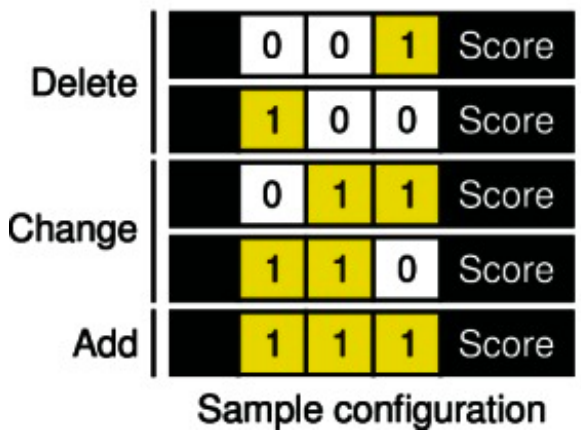


The shotgun stochastic search

Make some initial guess

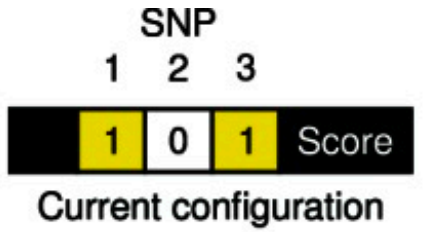


Search the neighborhood around the guess and calculate p* for each

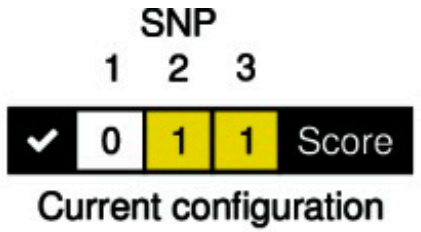


The shotgun stochastic search

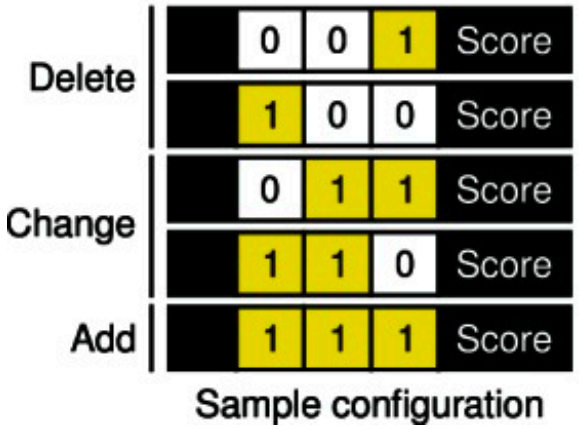
Make some initial guess



Sample a new configuration in the neighborhood with probability proportional to p*



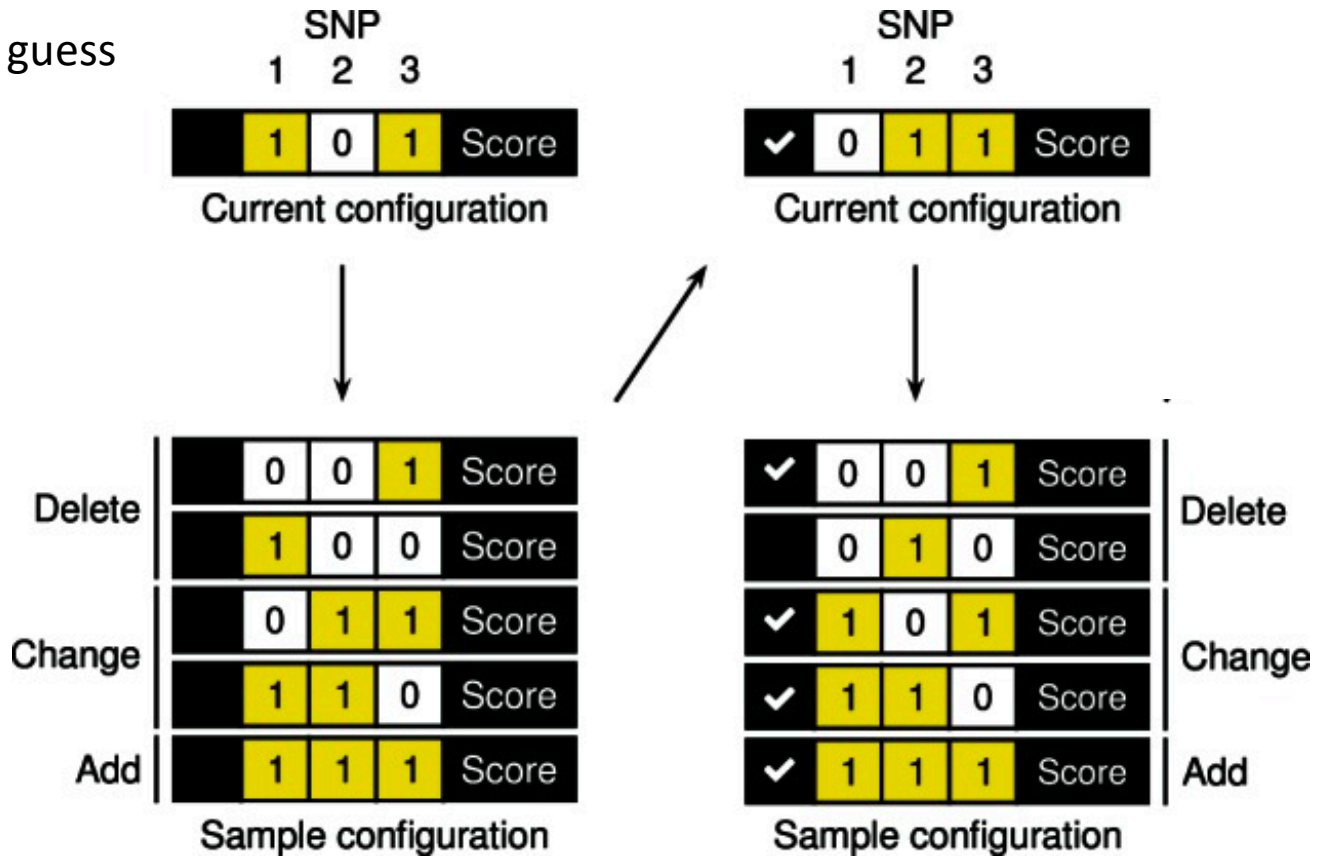
Search the neighborhood around the guess and calculate p* for each



The shotgun stochastic search

Make some initial guess

Sample a new configuration in the neighborhood with probability proportional to p*



Search the neighborhood around the guess and calculate p* for each

Keep searching, Save all the p*s as you go along.

The shotgun stochastic search

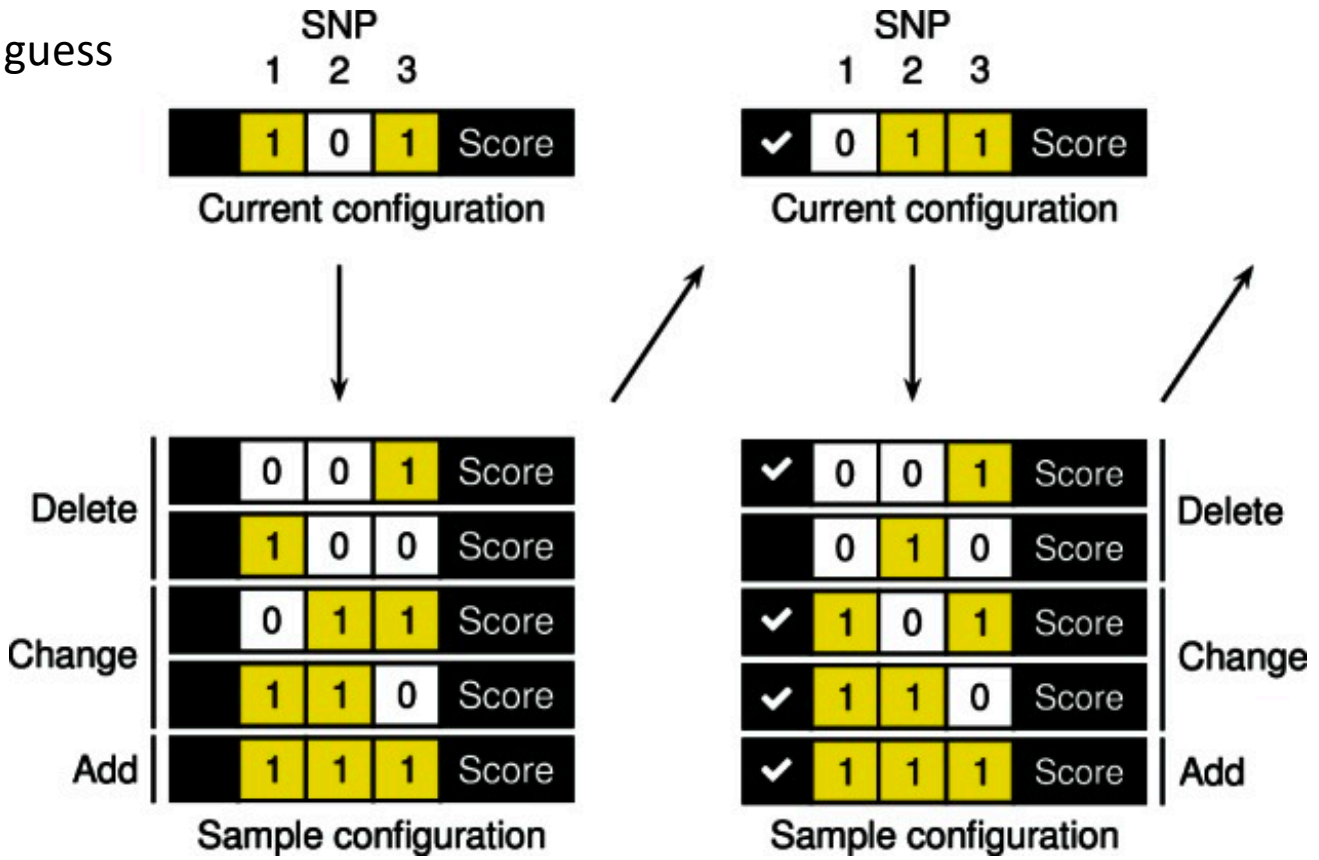
Make some initial guess

Search the neighborhood around the guess and calculate p* for each

Sample a new configuration in the neighborhood with probability proportional to p*

Keep searching, until the sum of all saved p*s stops increasing.

Keep searching, Save all the p*s as you go along.



FINEMAP outputs

FINEMAP outputs

$$BF(k > 0) = \frac{\sum_{\gamma \in \Gamma *} p * (\gamma | y, X)}{p(y | \gamma = \mathbf{0}, X)}$$

Bayes factor that there is at least one causal variant (i.e. evidence that there is any association at all)

FINEMAP outputs

$$BF(k > 0) = \frac{\sum_{\gamma \in \Gamma *} p * (\gamma | y, X)}{p(y | \gamma = \mathbf{0}, X)}$$

Bayes factor that there is at least one causal variant (i.e. evidence that there is any association at all)

$$p(k) = \frac{\sum_{\gamma \in \Gamma *; \sum_i \gamma_i = k} p * (\gamma | y, X)}{\sum_{\gamma \in \Gamma *} p * (\gamma | y, X)}$$

Probability that there are exactly k causal variants

FINEMAP outputs

$$BF(k>0) = \frac{\sum_{\gamma \in \Gamma *} p*(\gamma|y,X)}{p(y|\gamma = \mathbf{0}, X)}$$

Bayes factor that there is at least one causal variant (i.e. evidence that there is any association at all)

$$p(k) = \frac{\sum_{\gamma \in \Gamma *; \sum_i \gamma_i = k} p*(\gamma|y,X)}{\sum_{\gamma \in \Gamma *} p*(\gamma|y,X)}$$

Probability that there are exactly k causal variants

$$p(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}) = p^*(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}) \Big/ \sum_{\boldsymbol{\gamma} \in \Gamma^*} p^*(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}).$$

Posterior probabilities for all causal configurations in the search (by default, top 50k configurations are outputted).

FINEMAP outputs

$$BF(k>0) = \frac{\sum_{\gamma \in \Gamma *} p*(\gamma|y,X)}{p(y|\gamma = \mathbf{0}, X)}$$

Bayes factor that there is at least one causal variant (i.e. evidence that there is any association at all)

$$p(k) = \frac{\sum_{\gamma \in \Gamma *; \sum_i \gamma_i = k} p*(\gamma|y,X)}{\sum_{\gamma \in \Gamma *} p*(\gamma|y,X)}$$

Probability that there are exactly k causal variants

$$p(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}) = p^*(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}) \Big/ \sum_{\boldsymbol{\gamma} \in \Gamma^*} p^*(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}).$$

Posterior probabilities for all causal configurations in the search (by default, top 50k configurations are outputted).

$$p(\boldsymbol{\gamma}_\ell = 1|\mathbf{y},\boldsymbol{X}) = \sum_{\boldsymbol{\gamma} \in \Gamma^*} 1(\boldsymbol{\gamma}_\ell = 1) p(\boldsymbol{\gamma}|\mathbf{y},\boldsymbol{X}).$$

Posterior probabilities for each variant. Also uses these to calculate a credible set for each independent signal.

Back to the paper

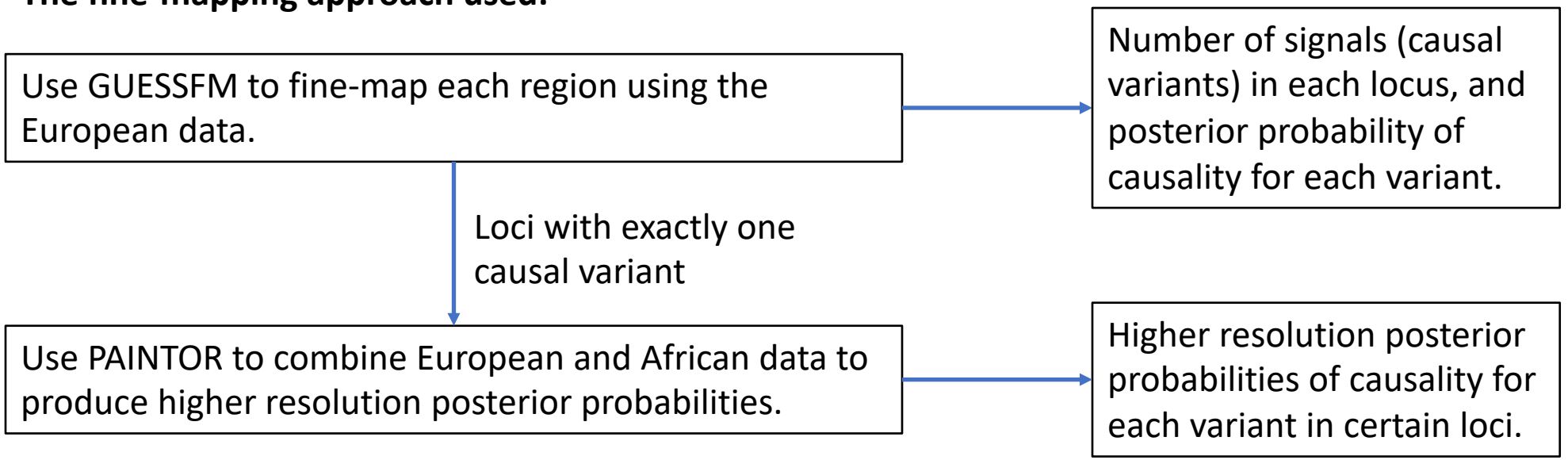
The fine-mapping approach used:

Use GUESSFM to fine-map each region using the European data.

Number of signals (causal variants) in each locus, and posterior probability of causality for each variant.

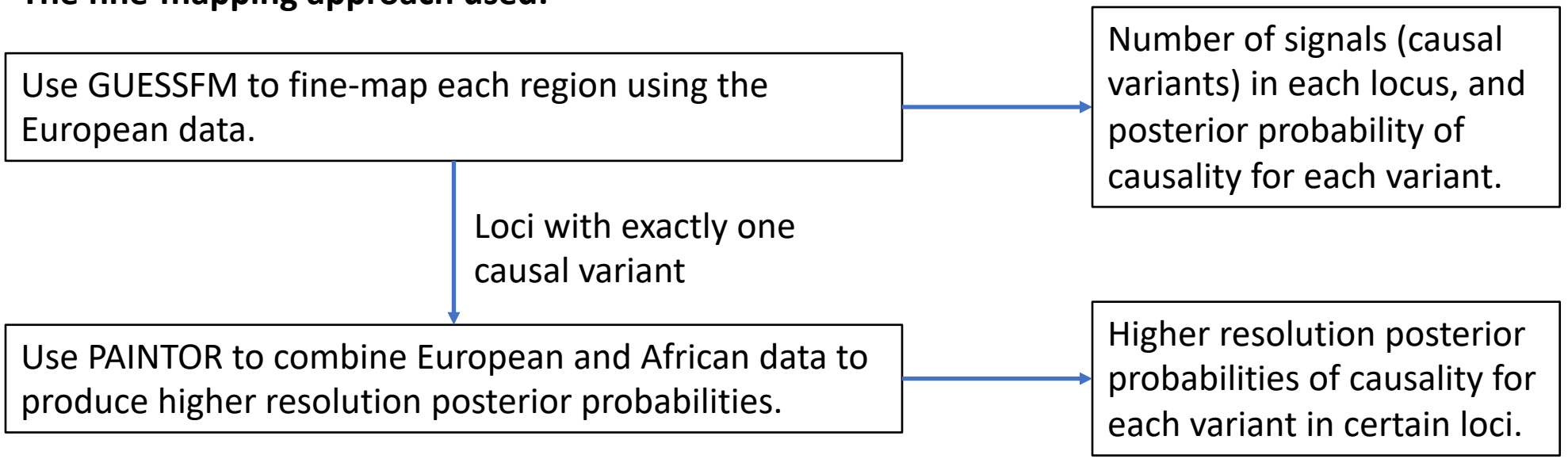
Back to the paper

The fine-mapping approach used:

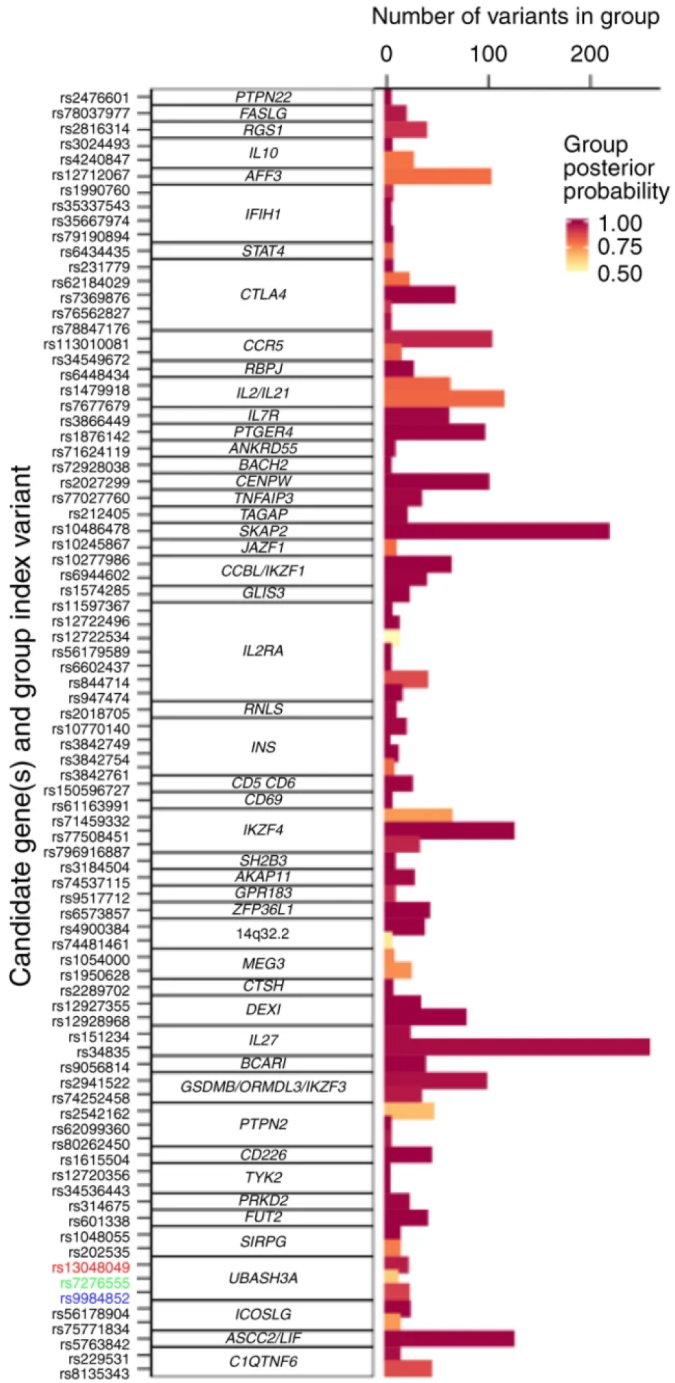


Back to the paper

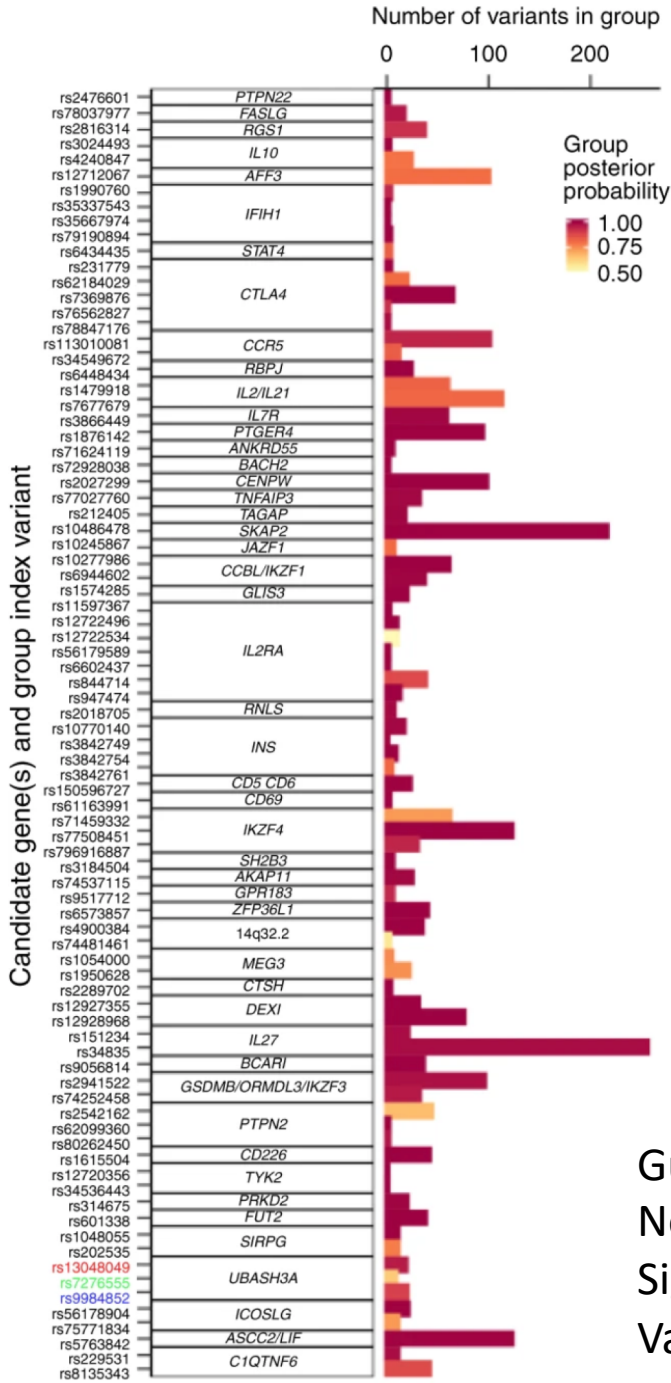
The fine-mapping approach used:



NOTE: PAINTOR is fine-mapping software that can combine data across ancestry groups, but it is unreliable when there are multiple causal variants.



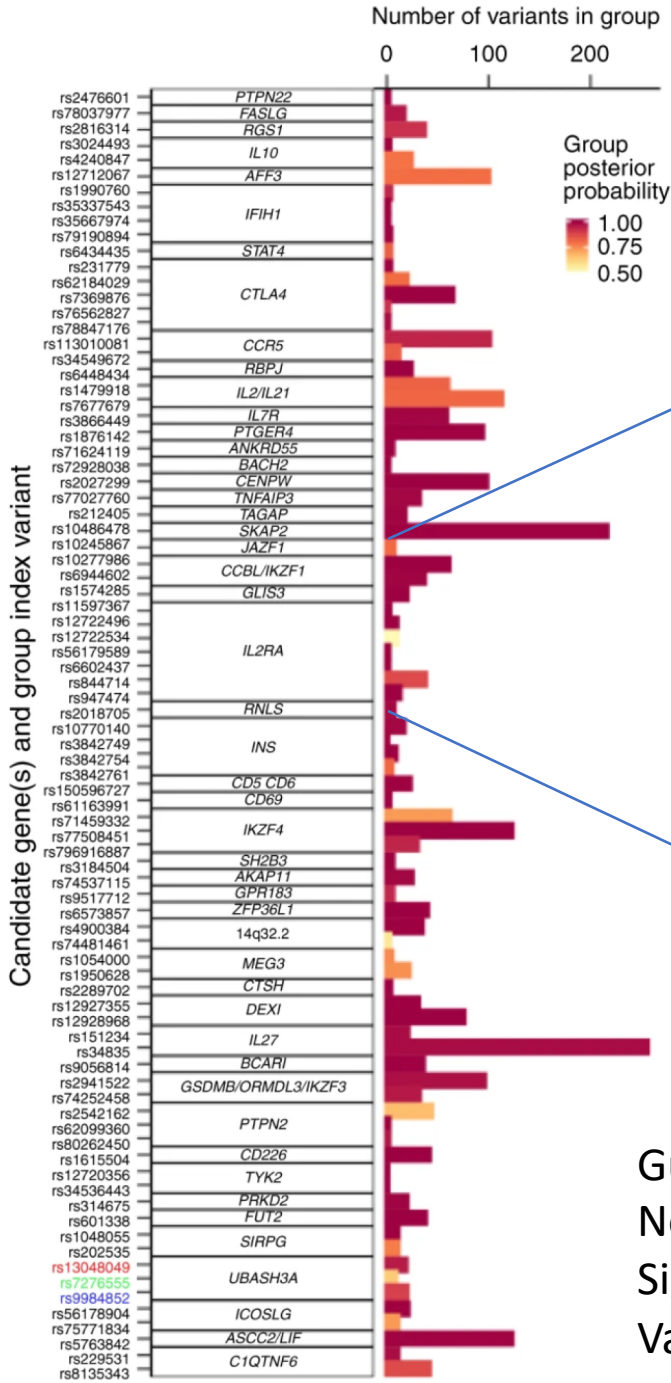
Identifying the number of causal variants at each locus

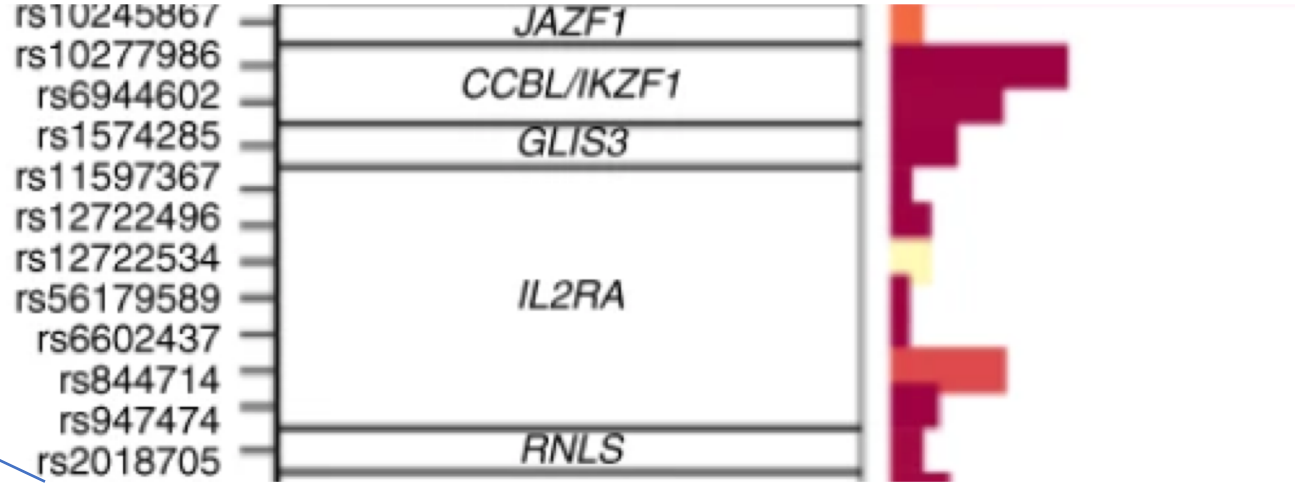


Identifying the number of causal variants at each locus

Guess FM output – all signals with posterior > 0.5.
Note difference between SIGNAL posterior and VARIANT posterior
Signal posterior == probability that there is an association here
Variant posterior = probability that this variant is a causal variant

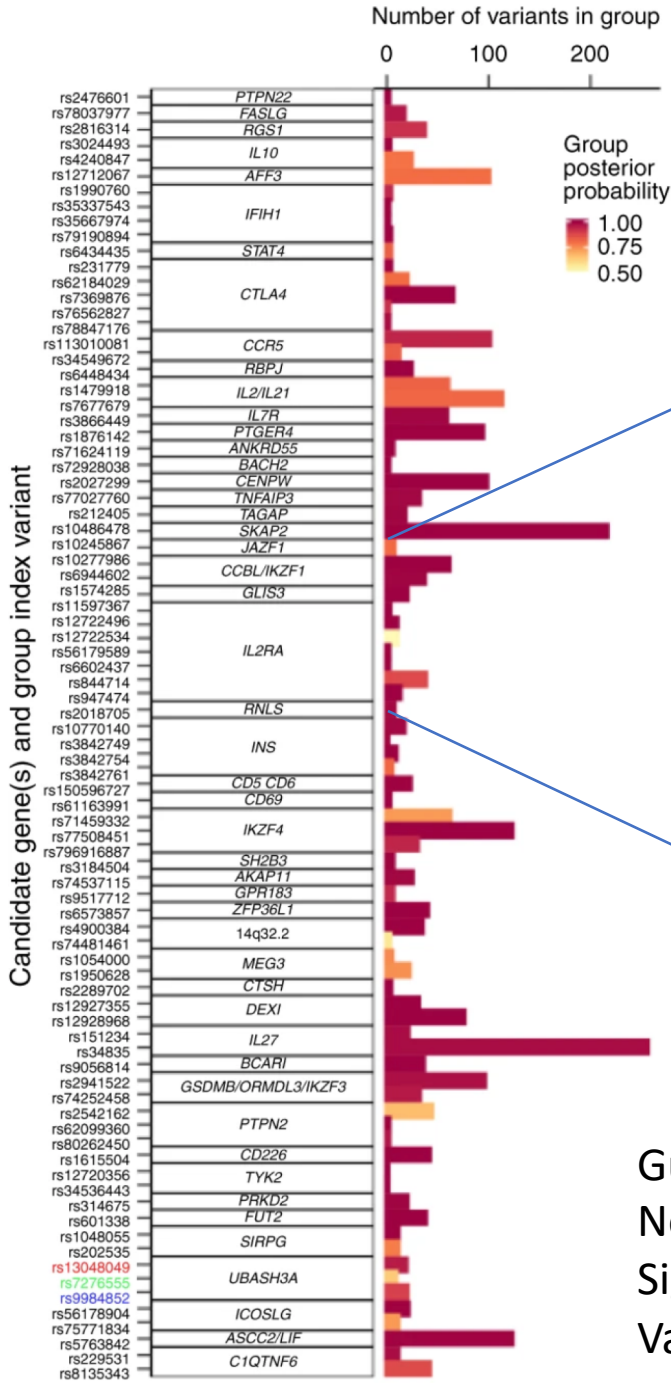
Identifying the number of causal variants at each locus

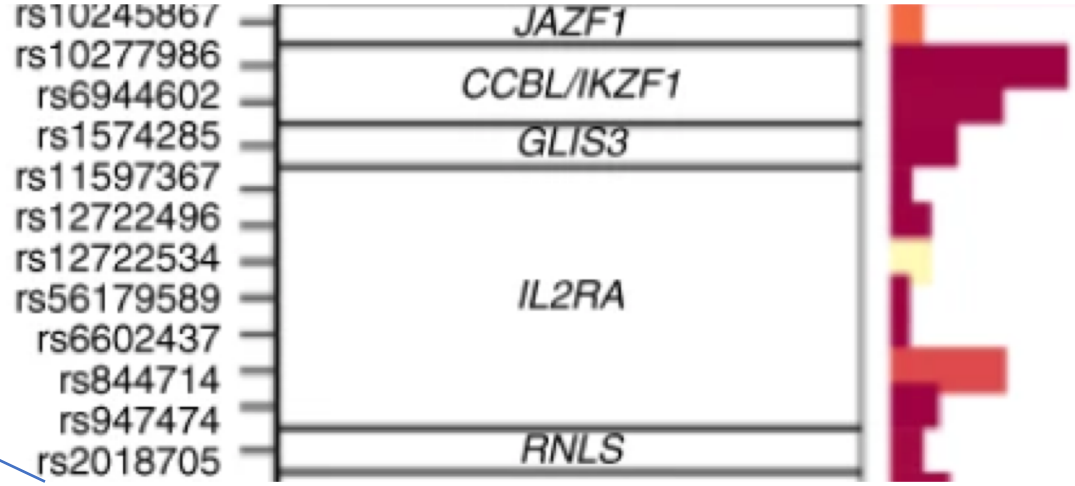




Guess FM output – all signals with posterior > 0.5.
Note difference between SIGNAL posterior and VARIANT posterior
Signal posterior == probability that there is an association here
Variant posterior = probability that this variant is a causal variant

Identifying the number of causal variants at each locus

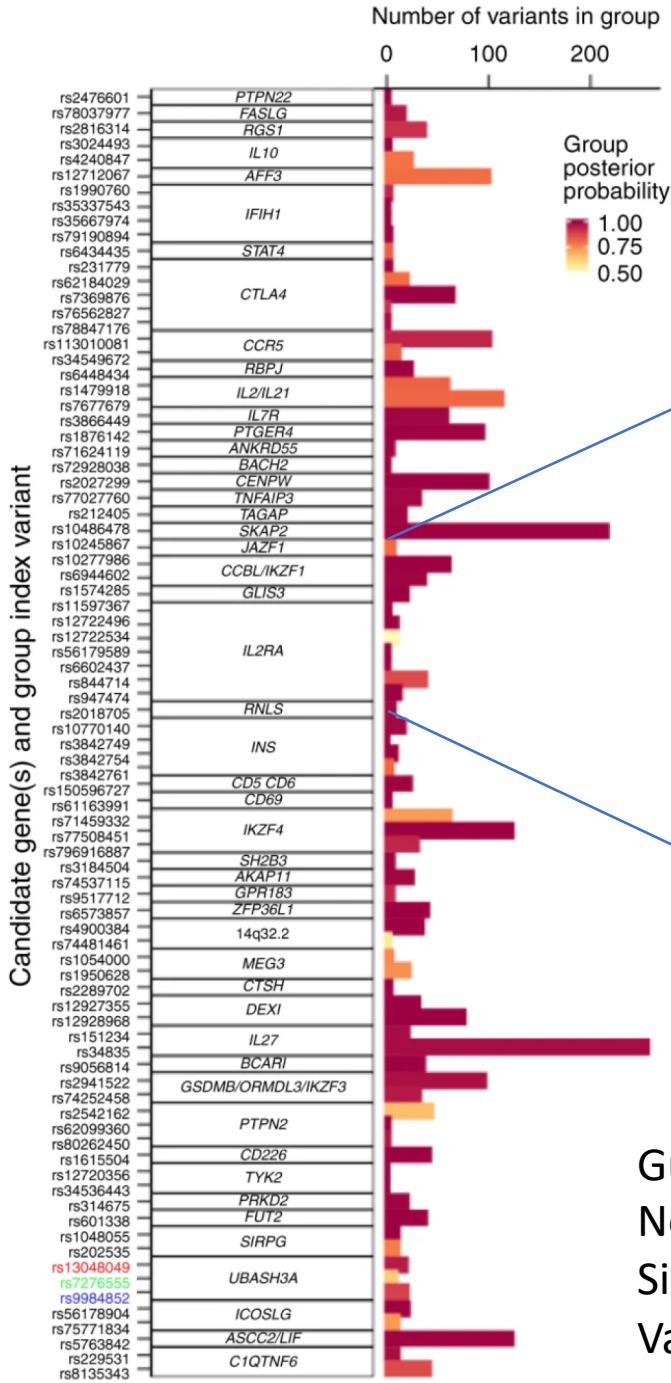


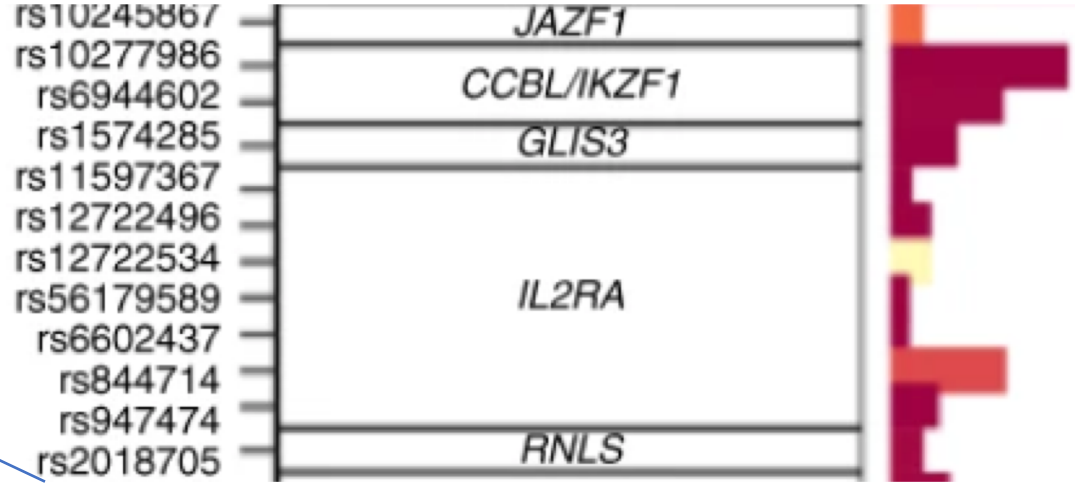


For JAZF1, GLIS3 and RNLS loci, the algorithm thinks there is only one causal variant.

Guess FM output – all signals with posterior > 0.5.
Note difference between SIGNAL posterior and VARIANT posterior
Signal posterior == probability that there is an association here
Variant posterior = probability that this variant is a causal variant

Identifying the number of causal variants at each locus



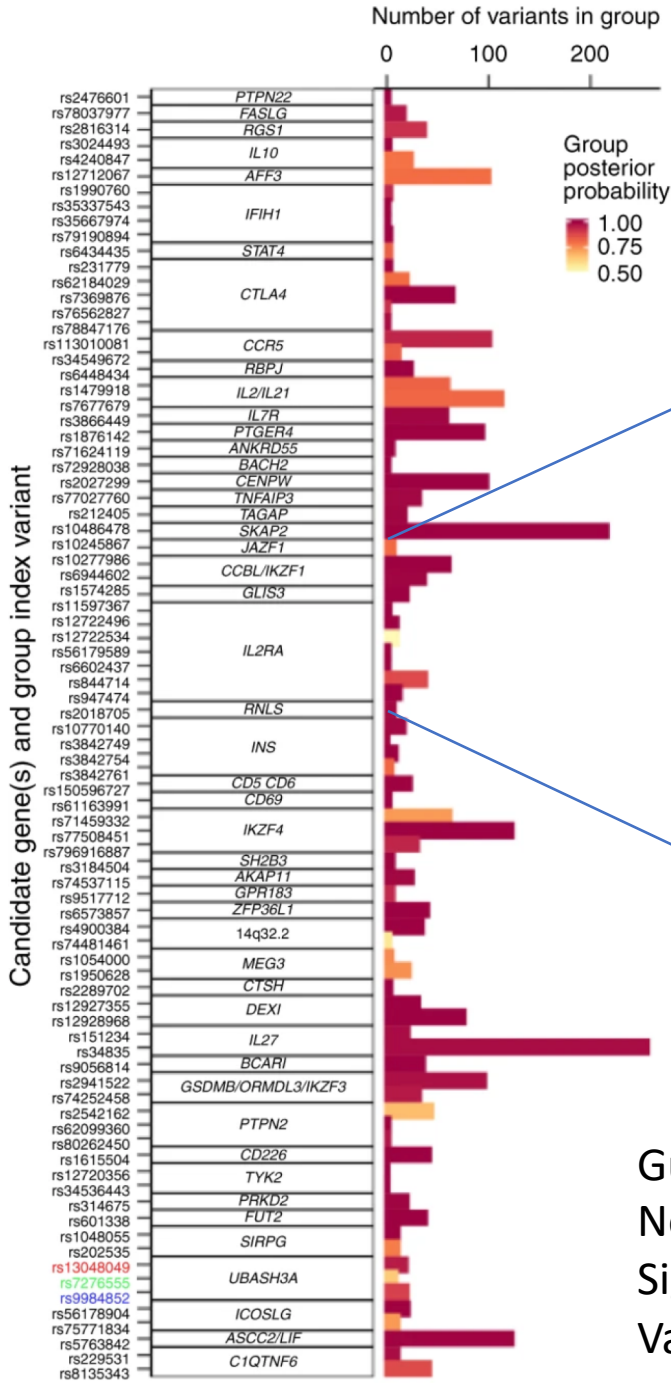


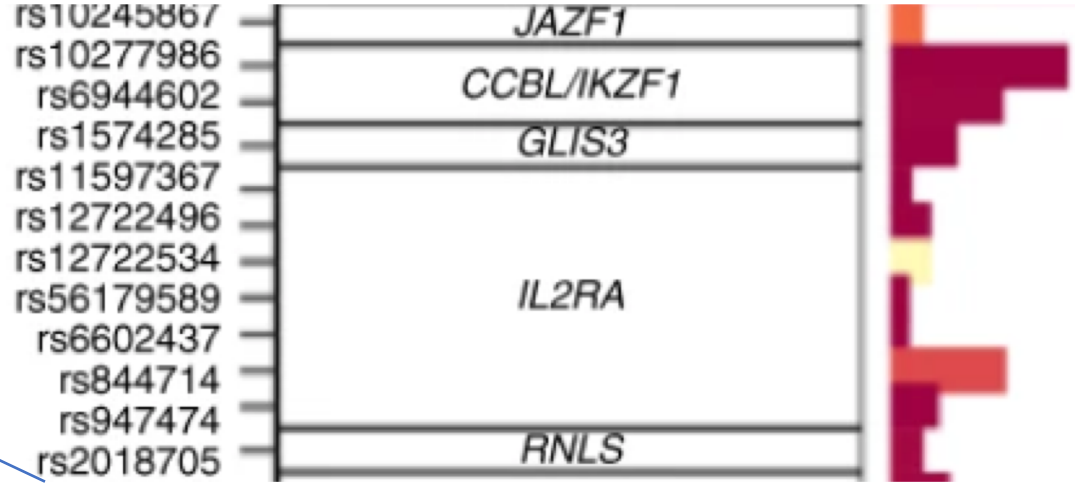
For JAZF1, GLIS3 and RNLS loci, the algorithm thinks there is only one causal variant.

For CCBL/IKZF1, it thinks there are two (high certainty).

Guess FM output – all signals with posterior > 0.5.
Note difference between SIGNAL posterior and VARIANT posterior
Signal posterior == probability that there is an association here
Variant posterior = probability that this variant is a causal variant

Identifying the number of causal variants at each locus





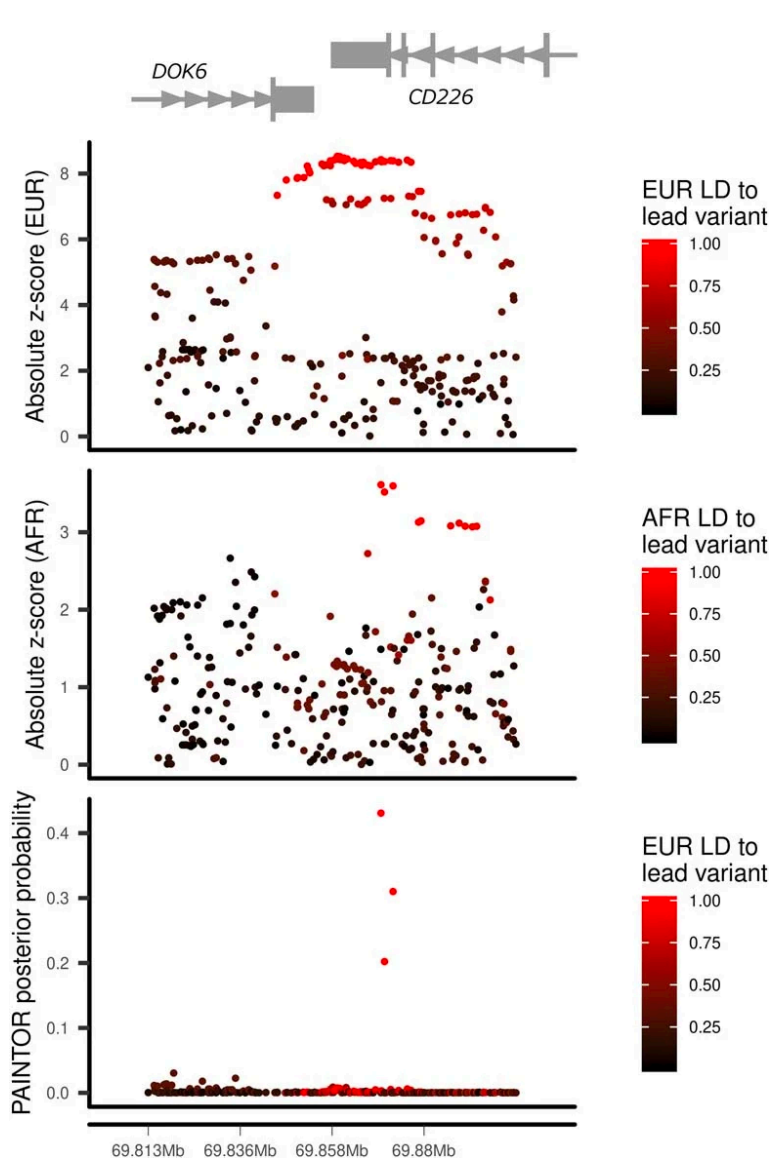
For JAZF1, GLIS3 and RNLS loci, the algorithm thinks there is only one causal variant.

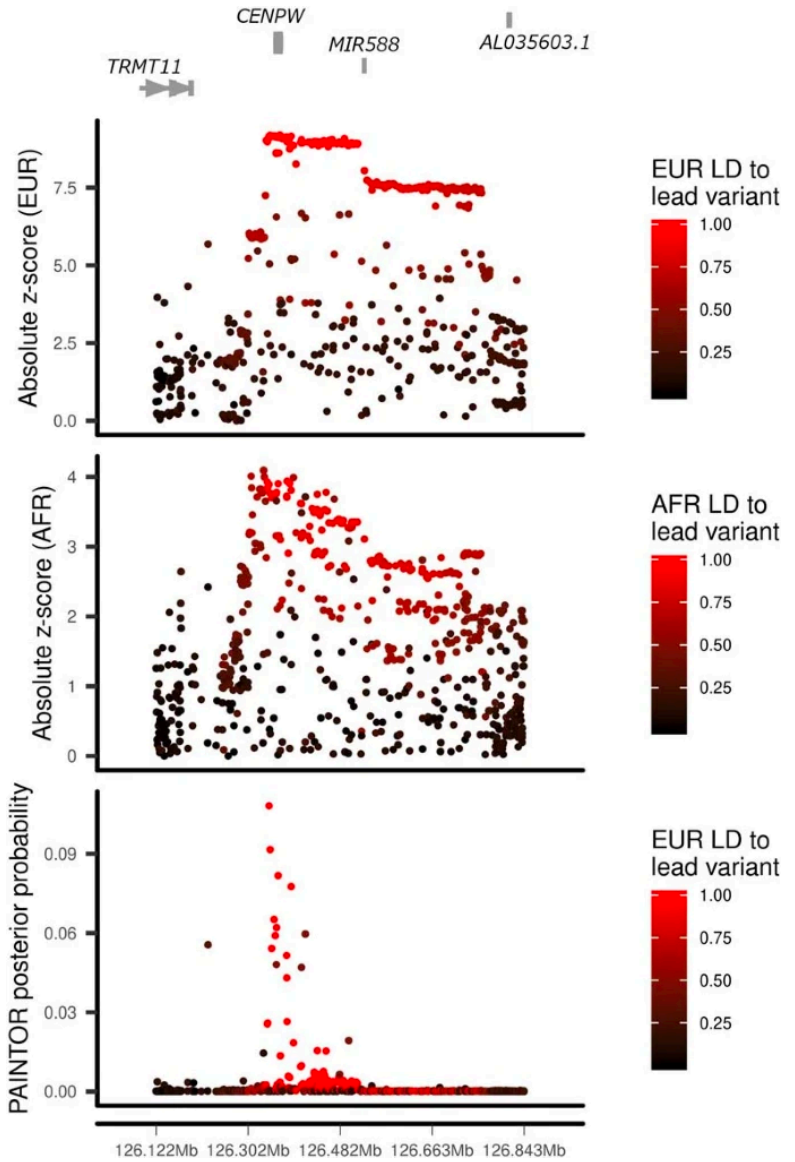
For CCBL/IKZF1, it thinks there are two (high certainty).

For IL2RA it thinks there are seven, though some are lower confidence (50-75%).

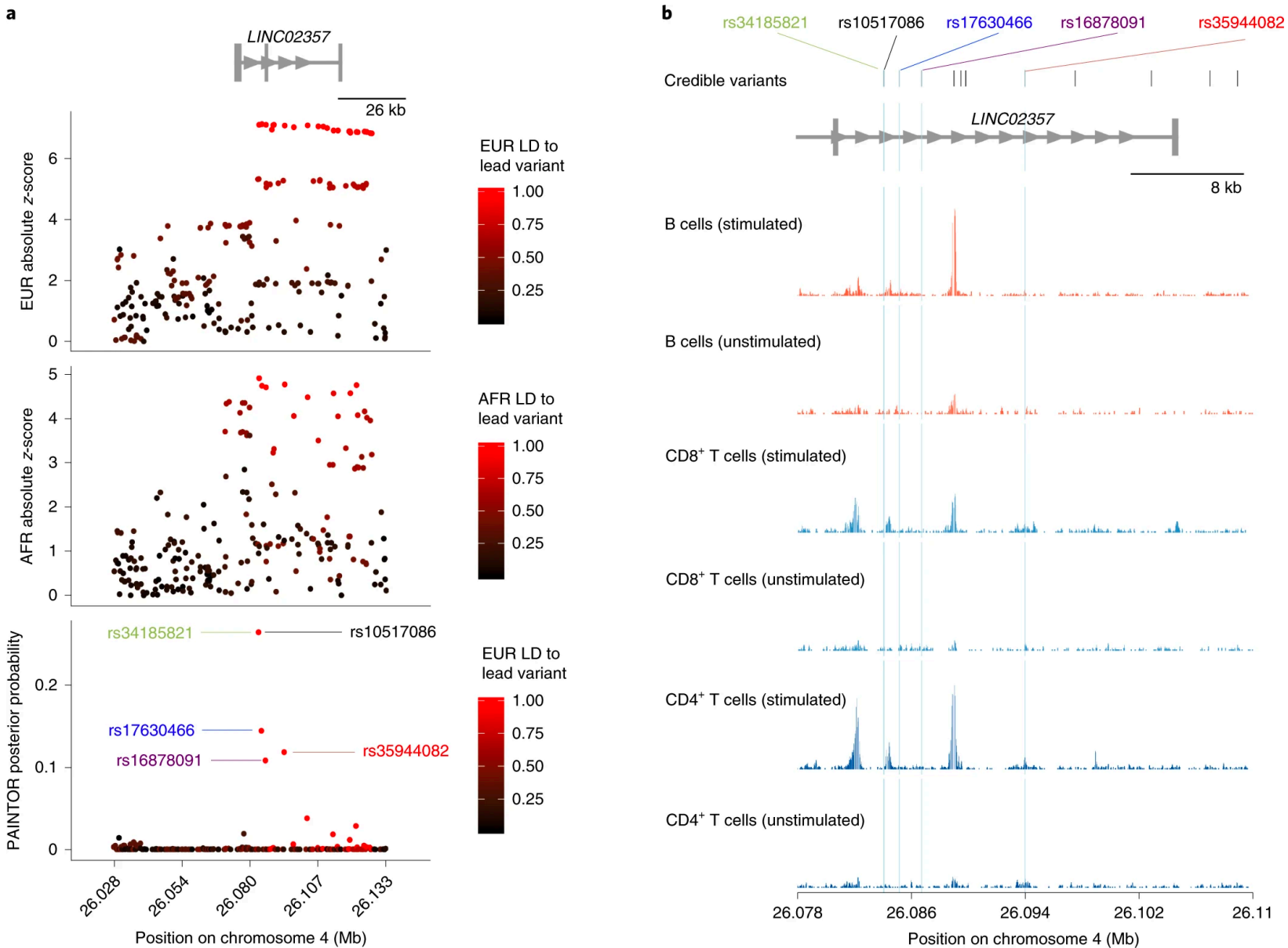
Guess FM output – all signals with posterior > 0.5.
Note difference between SIGNAL posterior and VARIANT posterior
Signal posterior == probability that there is an association here
Variant posterior = probability that this variant is a causal variant

Mapping causal variants using trans-ethnic data





Studying potential functions of causal variants



Discussion: following up causal variants

- How would you follow up a high-confidence causal variant in an experiment?
- How would this differ if you have 10 variant in the credible set, compared to 1?

Running these yourself

- Our practical will look at using FINEMAP to carry out fine-mapping using summary statistics.
- GUESSFM, which was used in the paper, is written in R and comes with some easy-to-use vignettes (which simulate their own example data):

https://chr1swallace.github.io/GUESSFM/

- We haven't discussed it at all, but the credible sets we have discussed are inherently Bayesian estimators. Anna Hutchinson has done some interesting work on putting Maller-style credible sets into a frequentist framework. Her R package (corrcoverage) has some well-written vignettes that talk you through this:

https://cran.r-project.org/web/packages/corrcoverage/

- The paper is good too:

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007829