

Analyses downstream of GWAS

SISG – Module 15

Dr Loic Yengo

l.yengo@imb.uq.edu.au

Institute for Molecular Bioscience
The University of Queensland

More information than we think...

- Initial motivation of GWAS was to map genetic variants associated with complex traits and disease...
- ...but
 - GWAS association are not (necessarily) causal (technology: SNP-chip and methods: marginal association)
 - Mapping capacity depends on sample size (power) yet SNP-based heritability is larger than variance explained by GWAS hits
 - Even if causal variants are identified, connection with underlying gene, biological pathway is not straightforward

So, what to do after your GWAS?

- Count how many distinct associations there are
- Fine-map your results (getting as close as possible to where “causal” variants are)
- Find what are the relevant genes?
- Find what are the relevant biological pathways?
- Prioritize tissues/cell-types of interest
- Prioritize drug targets?
- Write a Nature Genetics paper?
- Derive a genetic predictor for the trait of interest?

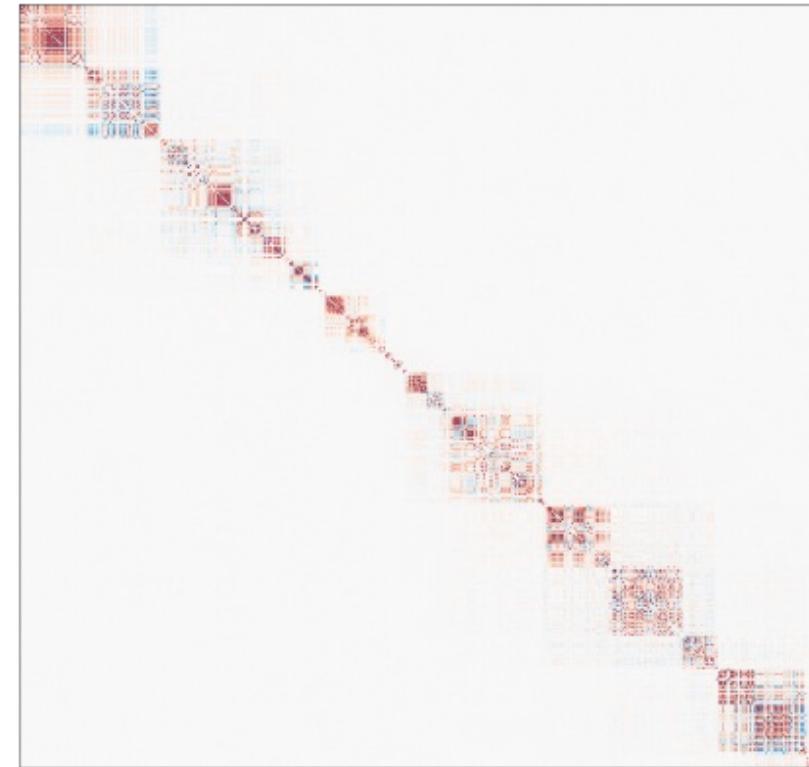
Outline of the lecture

- **Part 1: Single-variant resolution**
- **Part 2: Single-gene resolution**
- **Part 3: Gene sets / Pathway enrichment**

Part 1: Single-variant resolution

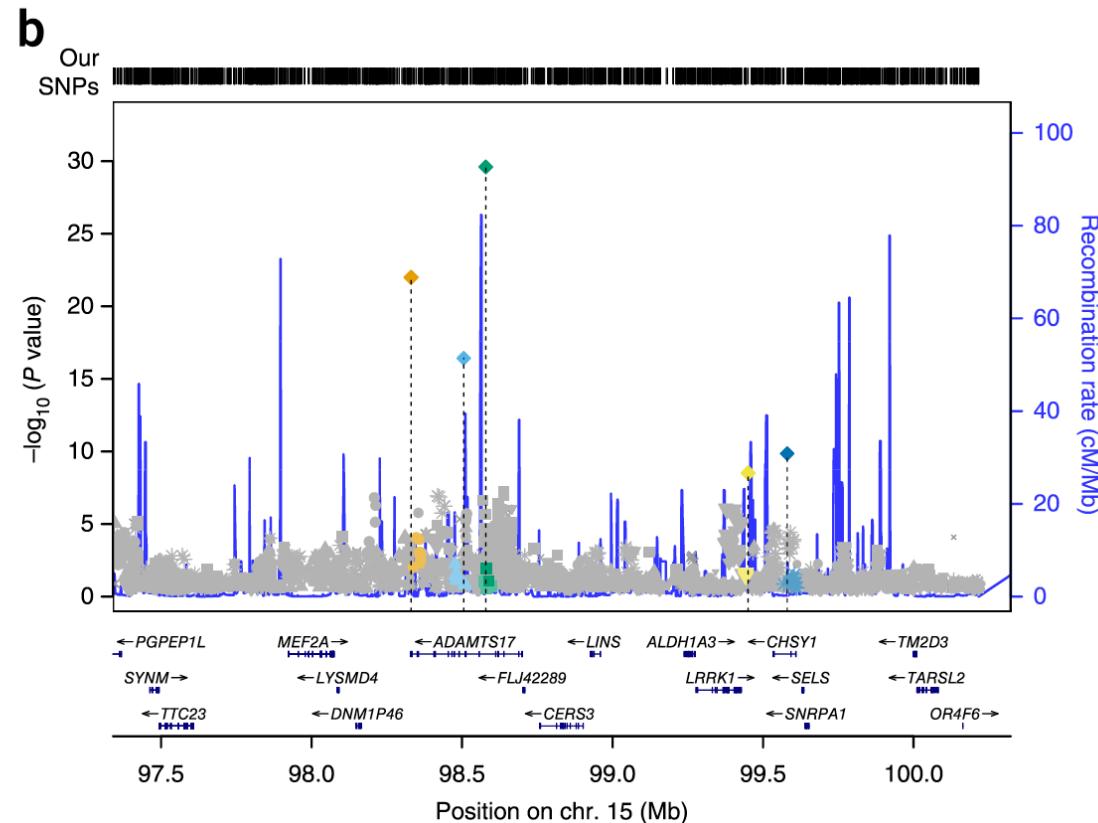
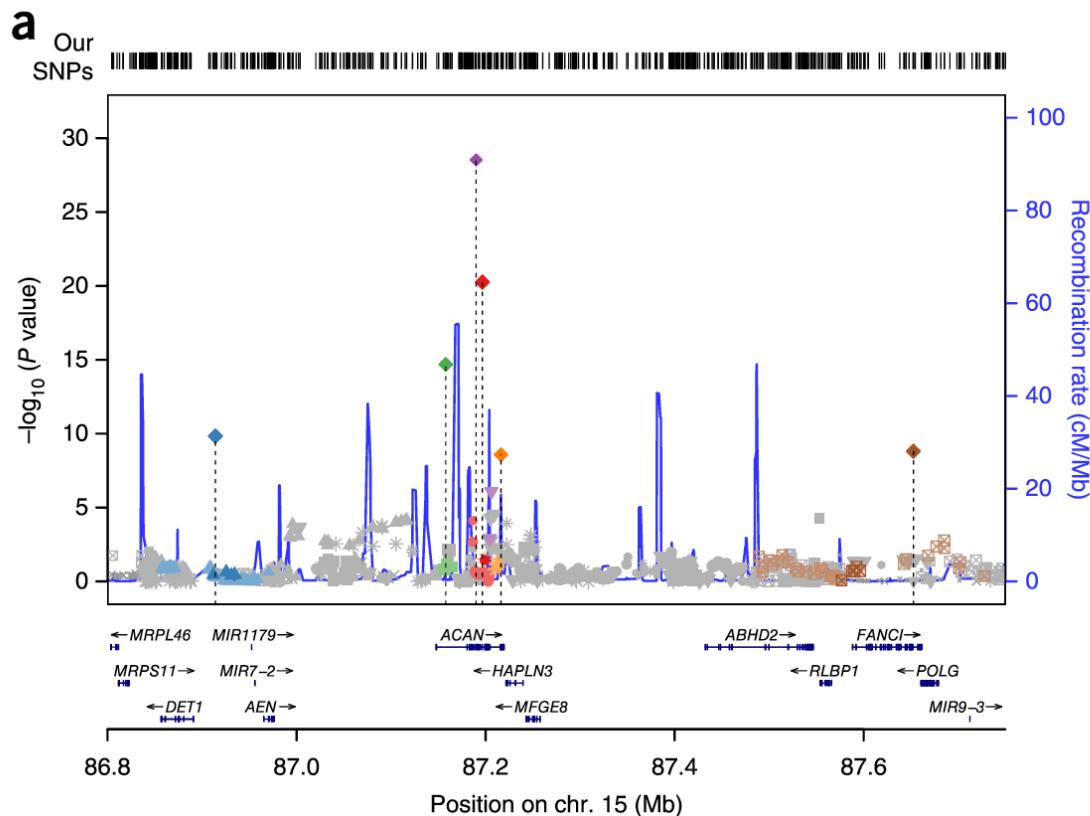
Counting the number distinct associations

- Classical algorithm is clumping (PLINK: Ben Voight algorithm)
- Given a significance threshold, select the most associated SNP within a certain window and discard all SNPs correlated with it above a certain squared correlation threshold



Linkage Disequilibrium (LD) correlation Matrix between SNPs (on a chromosome)

Counting the number distinct associations

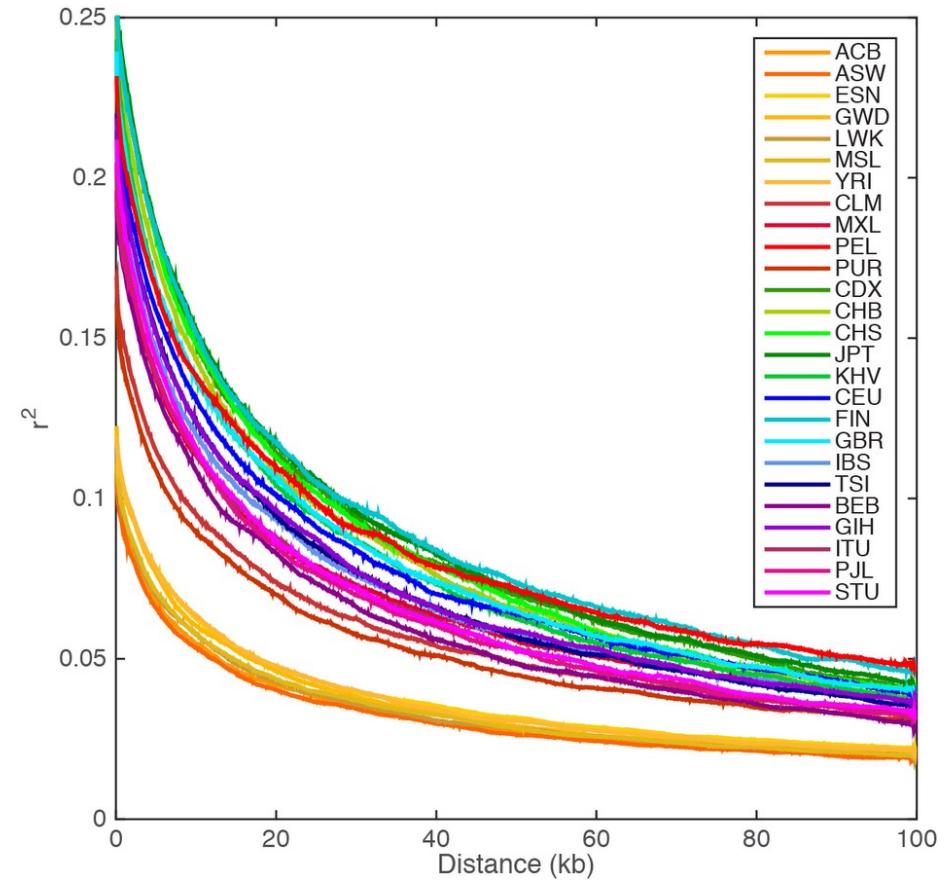


Wood et al. (2014)

Standard thresholds used on GWAS papers
(clumping $r^2 < 0.1$) within 1 Mb

Limitations

- Window size and LD are correlated but the relationship depends on ancestries (demographic history)
- $E[\chi^2] \approx 1 + Nr^2q^2$ so if a causal variant explains 0.25% of trait variance than if $N=300,000$ then $E[\chi^2] \approx 38$ for a SNP with an $r^2 = 0.05$ with the causal variant.



Auton et al. Nature (2015)
1000 Genomes paper

(Approximate) COnditional and JOint Analysis

Rationale

- 1) Mimic/Approximate stepwise variable selection
- 2) There is more variance explained within loci if SNPs partially in LD are independently associated with the trait

nature genetics

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature genetics](#) > [analyses](#) > [article](#)

Published: 18 March 2012

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

[Jian Yang](#), [Teresa Ferreira](#), [Andrew P Morris](#), [Sarah E Medland](#), [Genetic Investigation of ANthropometric Traits \(GIANT\) Consortium](#), [DIAbetes Genetics Replication And Meta-analysis \(DIAGRAM\) Consortium](#), [Pamela A F Madden](#), [Andrew C Heath](#), [Nicholas G Martin](#), [Grant W Montgomery](#), [Michael N Weedon](#), [Ruth J Loos](#), [Timothy M Frayling](#), [Mark I McCarthy](#), [Joel N Hirschhorn](#), [Michael E Goddard](#) & [Peter M Visscher](#)✉

COJO algorithm is implemented in the GCTA software (Yang et al. 2011)

Overview of the COJO algorithm

- 1 Start with a model with the most significant SNP in the single-SNP meta-analysis across the whole genome with P value below a cutoff P value, such as 5×10^{-8} .
- 2 For the t th step, calculate the P values of all the remaining SNPs conditional on the SNP(s) that have already been selected in the model. To avoid problems due to colinearity, if the squared multiple correlation between a SNP to be tested and the selected SNP(s) is larger than a cutoff value, such as 0.9, the conditional P value for that SNP will be set to 1.
- 3 Select the SNP with minimum conditional P value that is lower than the cutoff P value. However, if adding the new SNP causes new colinearity problems between any of the selected SNPs and the others, we drop the new SNP and repeat this process.
- 4 Fit all the selected SNPs jointly in a model and drop the SNP with the largest P value that is greater than the cutoff P value.
- 5 Repeat processes (2), (3) and (4) until no SNPs can be added or removed from the model.

Approximate Multivariate Linear Regression from GWAS summary statistics

Closed-form formula (ordinary least-squares)

$$\boldsymbol{\beta}_{joint} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = (\mathbf{N}^{-1}\mathbf{X}'\mathbf{X})^{-1}(\mathbf{N}^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{R}_{GWAS})^{-1}\mathbf{D}^{-1}\boldsymbol{\beta}_{GWAS}$$

\mathbf{X} is the matrix a $N \times M$ matrix of genotypes (at the focal locus) in the GWAS sample

\mathbf{y} is the vector of phenotypes for the N individuals in the GWAS

\mathbf{R}_{GWAS} is the correlation matrix between SNPs in the GWAS sample

$\boldsymbol{\beta}_{GWAS}$ is the vector marginal SNP effects

\mathbf{D} is a diagonal matrix: $D_j = \text{var}(X_j)$ is the variance of allele counts at SNP j in the GWAS sample.

Approximate Multivariate Linear Regression from GWAS summary statistics

Closed-form formula (ordinary least-squares)

$$\boldsymbol{\beta}_{joint} = (\mathbf{R}_{GWAS})^{-1} \mathbf{D}_{GWAS}^{-1} \boldsymbol{\beta}_{GWAS}$$

Issue: However, \mathbf{R}_{GWAS} and \mathbf{D}_{GWAS} may not always be available.

Solutions: approximate \mathbf{R}_{GWAS} and \mathbf{D}_{GWAS} from a sample of individuals with same genetic ancestries as the GWAS sample.

$$\boldsymbol{\beta}_{joint} \approx (\mathbf{R}_{REF})^{-1} \mathbf{D}_{REF}^{-1} \boldsymbol{\beta}_{GWAS}$$

This is a very fruitful idea utilized in many other summary statistics-based methods!

In GCTA language...

```
# Select multiple associated SNPs through a stepwise selection procedure
gcta64 --bfile test --cojo-p 5e-8 --cojo-file test.ma --cojo-slct --out test_chr1
```

--bfile: test (*.bed, bim, fam): specifies genotype data in PLINK format.

This is where LD will be calculated from

--cojo-p: significance threshold to select SNPs

--cojo-file: specifies GWAS summary statistic file

--cojo-slct: tells GCTA to run stepwise SNP selection

--out: prefix for the output files ([prefix].cojo.jma, [prefix].cojo.ldr, [prefix].cma, [prefix].log)

SNP	A1	A2	freq	b	se	p	N
rs1001	A	G	0.8493	0.0024	0.0055	0.6653	129850
rs1002	C	G	0.0306	0.0034	0.0115	0.7659	129799
rs1003	A	C	0.5128	0.0045	0.0038	0.2319	129830
...							

COJO for conditional analyses only

- Another application of COJO is for “single-step” conditional analysis
- You have a set of SNPs and want to know if your GWAS results are independent from these SNPs

In GCTA language...

```
gcta64 --bfile test \
--chr 1 \
--cojo-file test.ma
--cojo-cond cond.snplist
--out test_chr1
```

Challenges

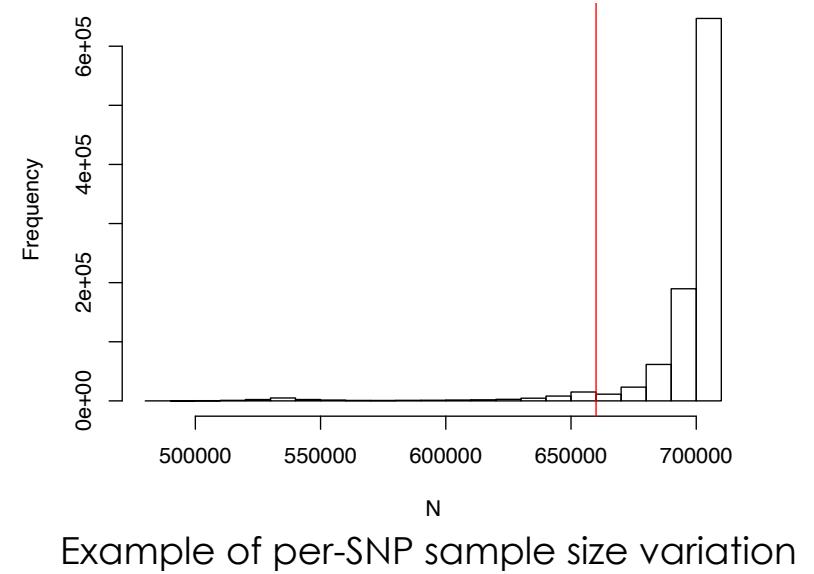
- 1) **Collinearity:** the correlation matrix (R) must be invertible
(--cojo-collinear: default is 0.9) – other solution (not in GCTA though)
is to shrink the LD matrix: $\mathbf{R}(s) = s\mathbf{R} + (1 - s)\mathbf{I}_M$.
- 1) Sample size of LD reference must be “large enough” (rule of thumb from GCTA, $N>4k$)
- 2) LD reference must match the correlation structure between SNPs
– otherwise, inflation or loss of power.

Mismatch between LD reference and GWAS summary statistics – causes

- 1) Ancestries mismatch (e.g., EUR LD panel for AFR GWAS or a Colombian LD panel for a Mexican GWAS)
- 2) Issues with GWAS summary statistics
 - 1) Truncated data (e.g., Wood et al. 2014)
 - 2) Variation in per-SNP sample size => SNPs effects are not all estimated on the same number of individuals (imputation)
 - 3) Allele frequencies may be that of reference (GIANT consortium)

Mismatch between LD reference and GWAS summary statistics – diagnostic

- 1) Compare allele frequencies between LD panel and GWAS
- 2) Filter out SNPs with too large differences (GCTA: --diff-freq)
- 3) Filter per-SNP sample size outliers
- 4) Use methods like DENTIST to detect LD inconsistencies



DENTIST, in brief

Test if the Z score of a given SNP is too far from its expected value given

1) Z-scores of neighboring SNPs

2) LD from a panel

Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors

[Wenhan Chen](#), [Yang Wu](#), [Zhili Zheng](#), [Ting Qi](#), [Peter M. Visscher](#), [Zhihong Zhu](#) & [Jian Yang](#)

[Nature Communications](#) 12, Article number: 7117 (2021) | [Cite this article](#)

4583 Accesses | 1 Citations | 9 Altmetric | [Metrics](#)

Application / Similar idea used in Kanai et al.

Meta-analysis fine-mapping is often miscalibrated at single-variant resolution

Masahiro Kanai^{1,2,3,4,5,*}, Roy Elzur^{1,2,3}, Wei Zhou^{1,2,3}, Global Biobank Meta-analysis Initiative, Mark J Daly^{1,2,3,6}, Hilary K Finucane^{1,2,3,*}

¹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA,

²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, ³Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, ⁴Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, ⁵Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan, ⁶Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

* Corresponding authors: Masahiro Kanai (mkanai@broadinstitute.org) and Hilary K Finucane (finucane@broadinstitute.org)

<https://www.medrxiv.org/content/10.1101/2022.03.16.22272457v1.full.pdf>

Take-home messages

- QC summary statistics is important...for COJO and any other downstream analyses.
- Some methods may require an advanced QC but you'd need to know what assumptions are violated (expert).
- Methods using LD references (e.g., COJO) face collinearity issues. Solutions include constrain which SNP enters the model (ala GCTA) or shrinkage (a la GCTB)

(Bayesian) Fine-mapping

Aim: Identify a set of SNPs (a.k.a credible set) with a given probability to contain causal variants,

Popular methods include

- FINEMAP
- SUSIE
- CAVIAR
- PAINTOR
- GCTA-COJO?

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$f(\boldsymbol{\beta} | \text{Summary data}) \propto f(\text{Summary data} | \boldsymbol{\beta}) \times f(\boldsymbol{\beta})$$

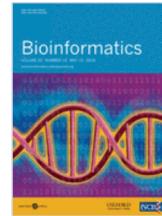
Single Causal variant model

- m SNPs in the region to fine-map
- Prior = each SNP has the same probability to be causal
- Posterior

$$P(C_j | Z_1, \dots, Z_m) = \frac{\exp\left(\frac{z_j^2}{2}\right)}{\sum_k^m \exp\left(\frac{z_k^2}{2}\right)}$$

FINEMAP overview

<http://www.christianbenner.com/>



Volume 32, Issue 10
15 May 2016

FINEMAP: efficient variable selection using summary data from genome-wide association studies

Christian Benner  Chris C.A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, Matti Pirinen  Author Notes

Bioinformatics, Volume 32, Issue 10, 15 May 2016, Pages 1493–1501, <https://doi.org/10.1093/bioinformatics/btw018>

Published: 14 January 2016 Article history ▾

- Requires the user to specifies the expected maximum number (k) of causal variants in the region
- Prior distribution explicitly considers all “causal configurations” –
 $C(m,k)=m!/[k! (m-k)!]$
(e.g., $m=100, k=3 \Rightarrow C(m,k) = 161,700$)



SuSiE: Sum of Single Effects

A simple new approach to variable selection in regression, with application to genetic fine-mapping

Gao Wang¹ Abhishek Sarkar¹ Peter Carbonetto^{1,2}
Matthew Stephens^{1,3}

- SuSiE = Bayesian stepwise selection (not best selected ala COJO but samples from a distribution)
- Algorithm is fast
- Seems a bit more robust than FINEMAP

Challenges

- 1) Regions must be well defined
(no consensus – depends on power)

- 1) FINEMAP is sensitive to the maximum
of causal variants specified

- 2) In practice people use different
methods and focus on consensus
results (e.g., FINEMAP + SuSiE)



THE PREPRINT SERVER FOR HEALTH SCIENCES



BMJ Yale

Insights from complex trait fine-mapping across diverse populations

DOI: Masahiro Kanai, Jacob C Ulirsch, Juha Karjalainen, Mitja Kurki, Konrad J Karczewski, Eric Fauman, Qingbo S Wang, Hannah Jacobs, François Aguet, Kristin G Ardlie, Nurlan Kerimov, Kaur Alasoo, Christian Benner, Kazuyoshi Ishigaki, Saori Sakaue, Steven Reilly, The BioBank Japan Project, FinnGen, Yoichiro Kamatani, Koichi Matsuda, Aarno Palotie, Benjamin M Neale, Ryan Tewhey, Pardis C Sabeti, Yukinori Okada, Mark J Daly, Hilary K Finucane

doi: <https://doi.org/10.1101/2021.09.03.21262975>

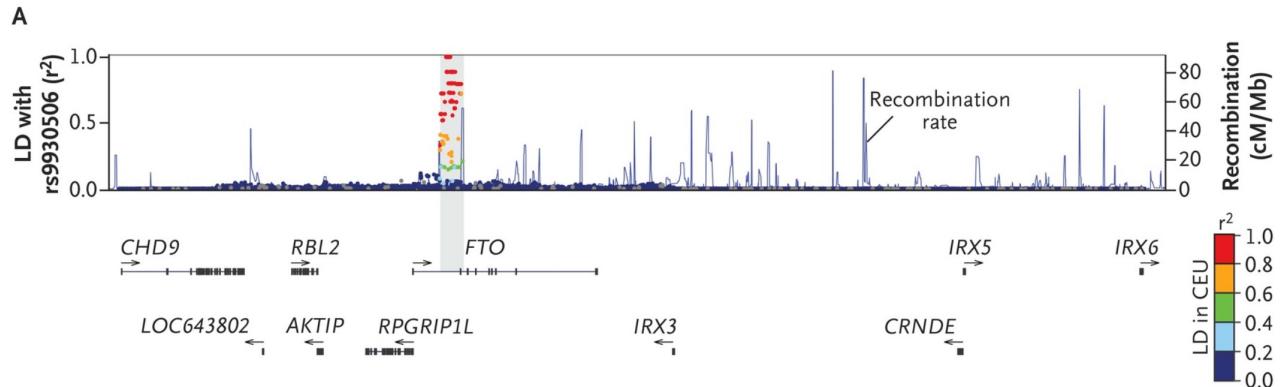
Part 2: Single-gene resolution

SNP to Gene mapping

- Can GWAS identify genes? If so, then how?
- **Approach 1** (positional mapping): relevant genes must be nearby...
- **Approach 2** (MR/Colocalization): SNP affecting trait/disease may also affect gene/protein expression

Approach 1 (positional mapping)

- Basic idea is to prioritize genes in the close vicinity of top GWAS hits
- Could also derive a gene-based test by aggregating association test-statistic at the level of genes



Claussnitzer et al. NEJM (2015)

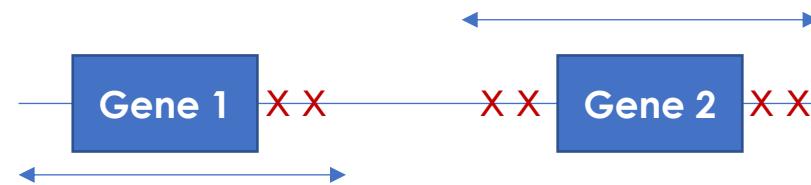
Gene-based tests

- **Step 1:** allocate SNPs to Genes
(position mapping, e.g., within 100 kb of TSS)
- **Step 2:** aggregate Z-scores
(sum of squared Z-scores*)

VEGAS (PLINK)

fastBAT (GCTA)

MAGMA



$$T = \sum_{j=1}^m z_j^2$$

*Main differences are how p-values of the aggregated statistic are calculated

Approach 2 (MR/Colocalization)

- Mendelian Randomization (MR) answers the question

“Does gene expression cause traits?”

- Colocalization answers the question
- “Are gene and trait expressions caused by the same variants?”

AJHG

Volume 109, Issue 5, 5 May 2022, Pages 767-782

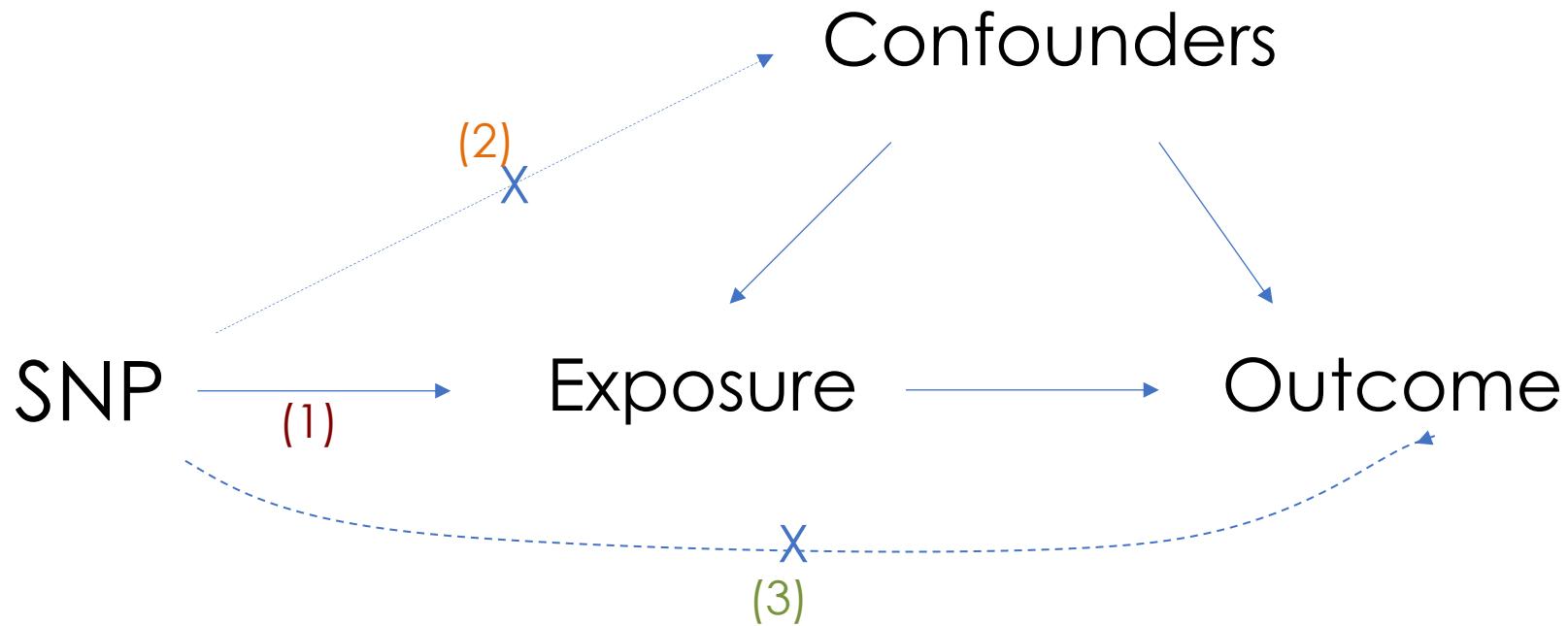


Review

Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches

Verena Zuber ^{1, 2, 3, 12}, Nastasiya F. Grinberg ^{4, 12}, Dipender Gill ^{1, 5, 6, 7}, Ichcha Manipur ^{8, 9}, Eric A.W. Slob ¹⁰, Ashish Patel ¹⁰, Chris Wallace ^{8, 9, 10}, Stephen Burgess ^{10, 11}  

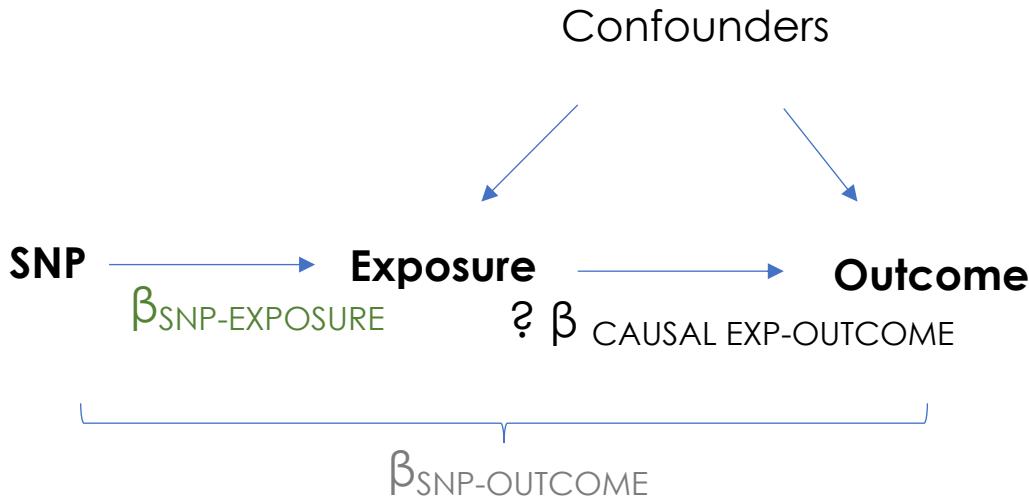
Mendelian Randomization: 3 Core Assumptions



- (1) SNP is associated with the exposure
- (2) SNP is NOT associated with confounding variables
- (3) SNP ONLY associated with outcome through the exposure

Calculating Causal Effect Estimates

Credit: Slides from Prof. D. Evans (UQ)

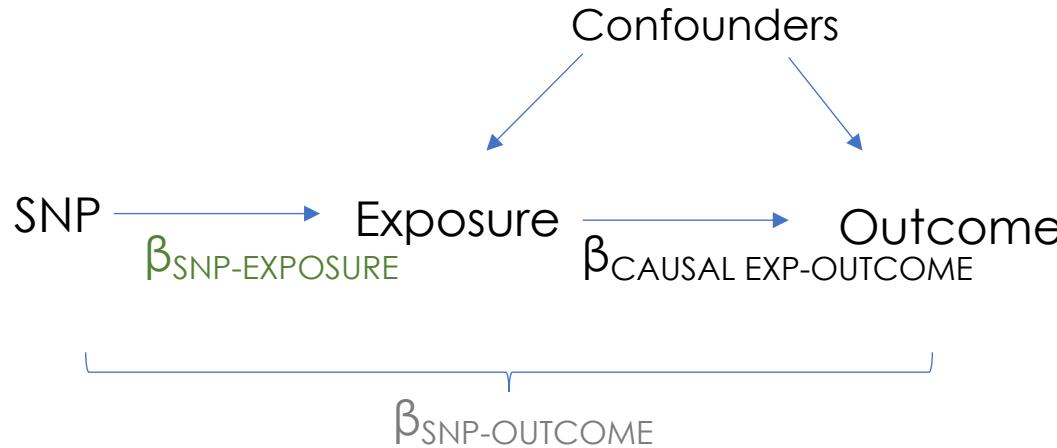


After SNP identified robustly associated with exposure of interest:

- Wald Estimator
- Two-stage least-squares (TSLS) regression

Calculating Causal Effect Estimates

Credit: Slides from Prof. D. Evans (UQ)



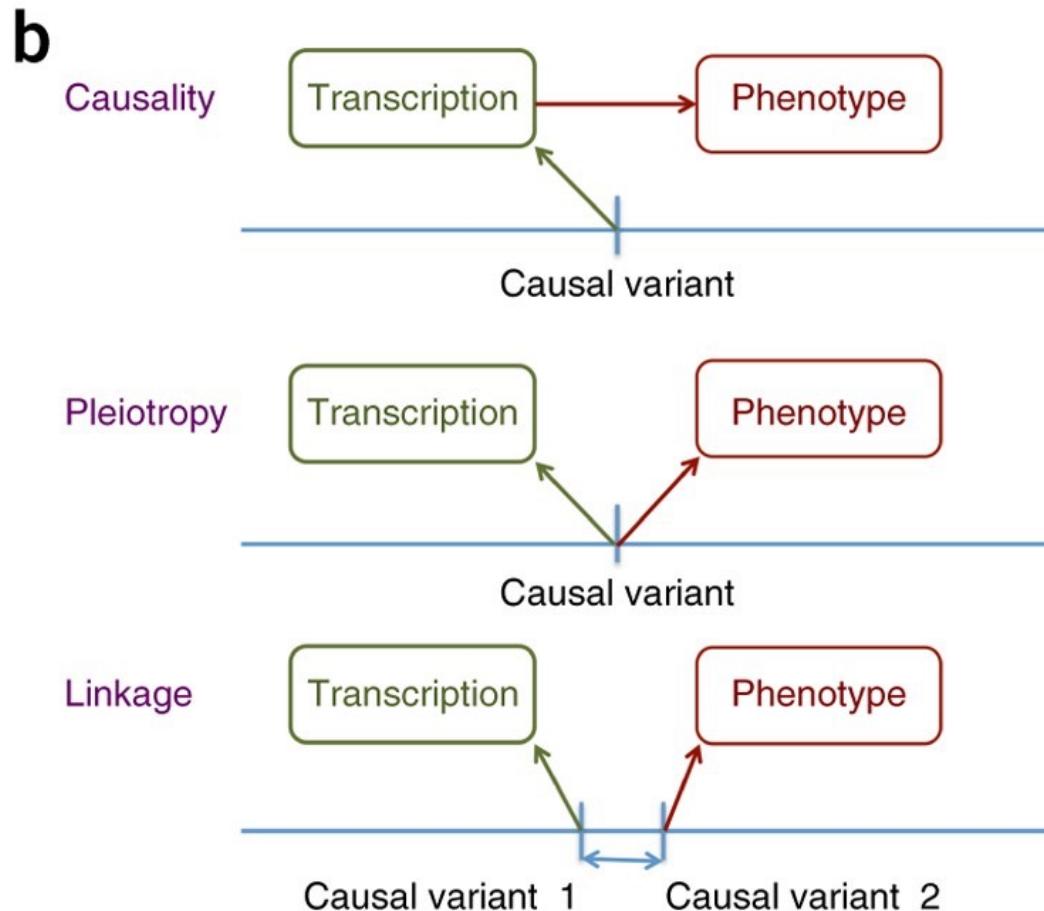
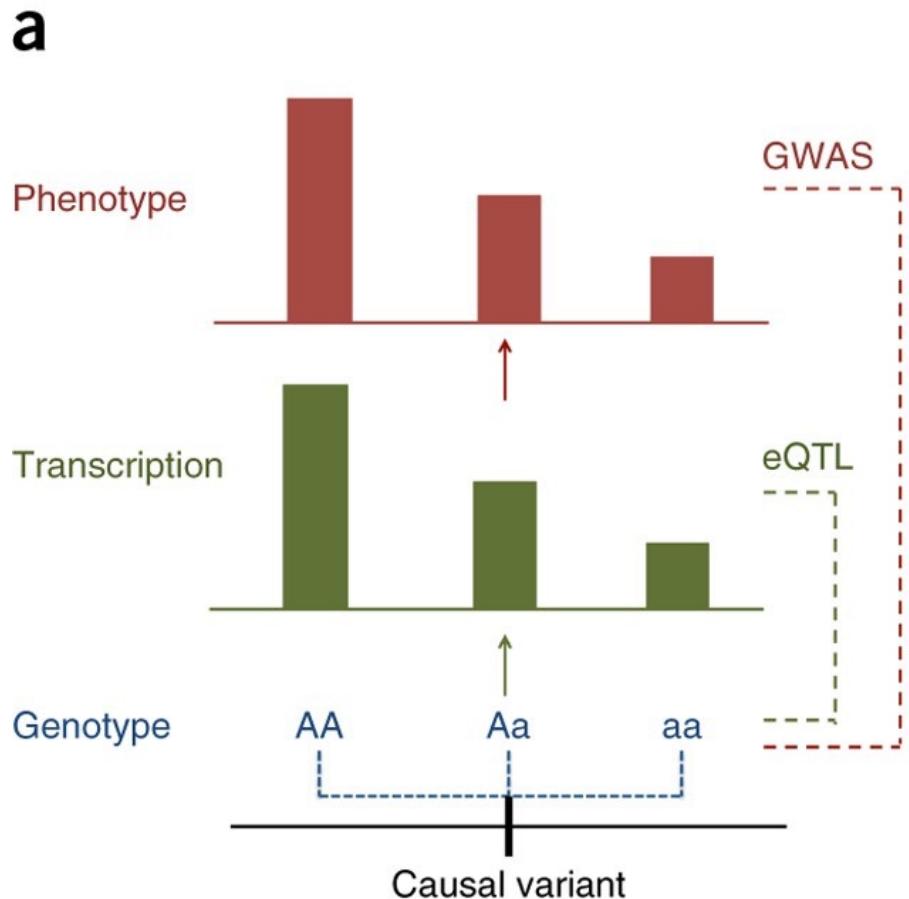
Causal effect by Wald Estimator* :

$$\frac{\beta_{SNP-OUTCOME}}{\beta_{SNP-EXPOSURE}}$$

$$\beta_{SNP-OUTCOME} = \beta_{CAUSAL EXP-OUTCOME} \times \beta_{SNP-EXPOSURE}$$

*Can be used in different samples ("Two sample MR")

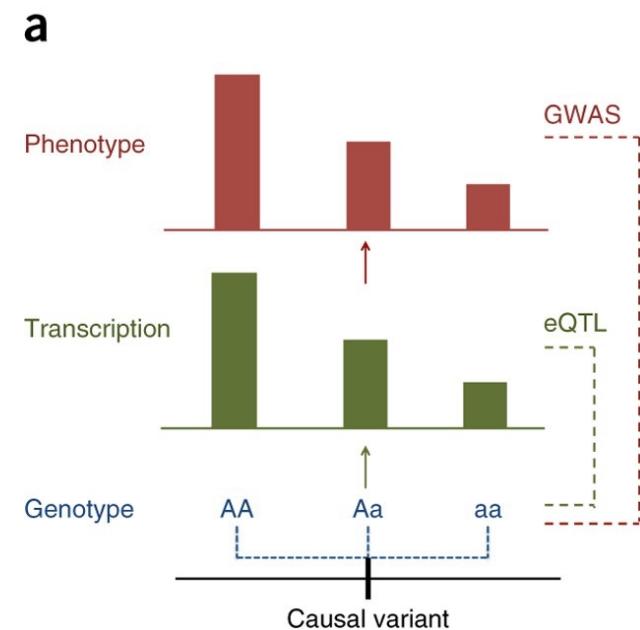
Summary-data based Mendelian Randomization (SMR) – Zhu et al. (2016)



GCTA-SMR language...

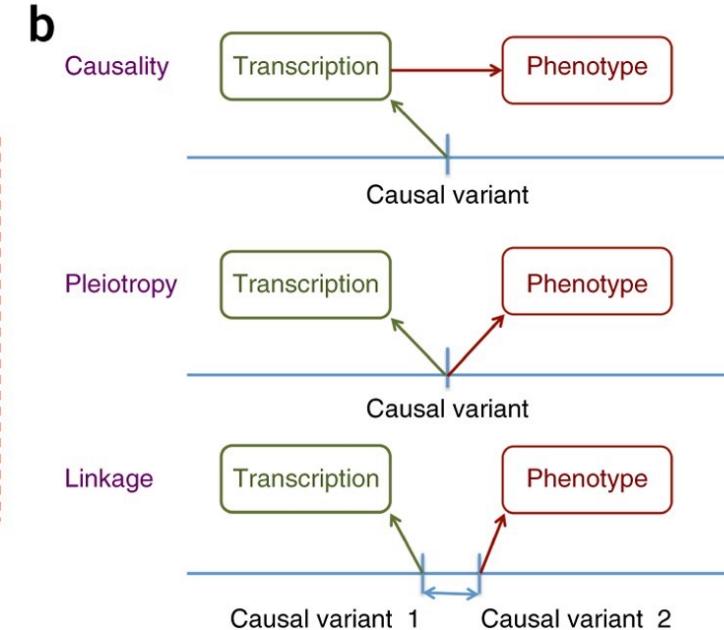
```
./smr \
--bfile mydata \
--gwas-summary mygwas.ma \
--beqtl-summary myeqtl \
--out mysmr
```

SNP	A1	A2	freq	b	se	p	n
rs1001	A	G	0.8493	0.0024	0.0055	0.6653	129850
rs1002	C	G	0.03606	0.0034	0.0115	0.7659	129799
rs1003	A	C	0.5128	0.045	0.038	0.2319	129830
.....							



eQTL, mQTL, pQTL, caQTL, etc.
Can be downloaded from the SMR website

Same format as for COJO

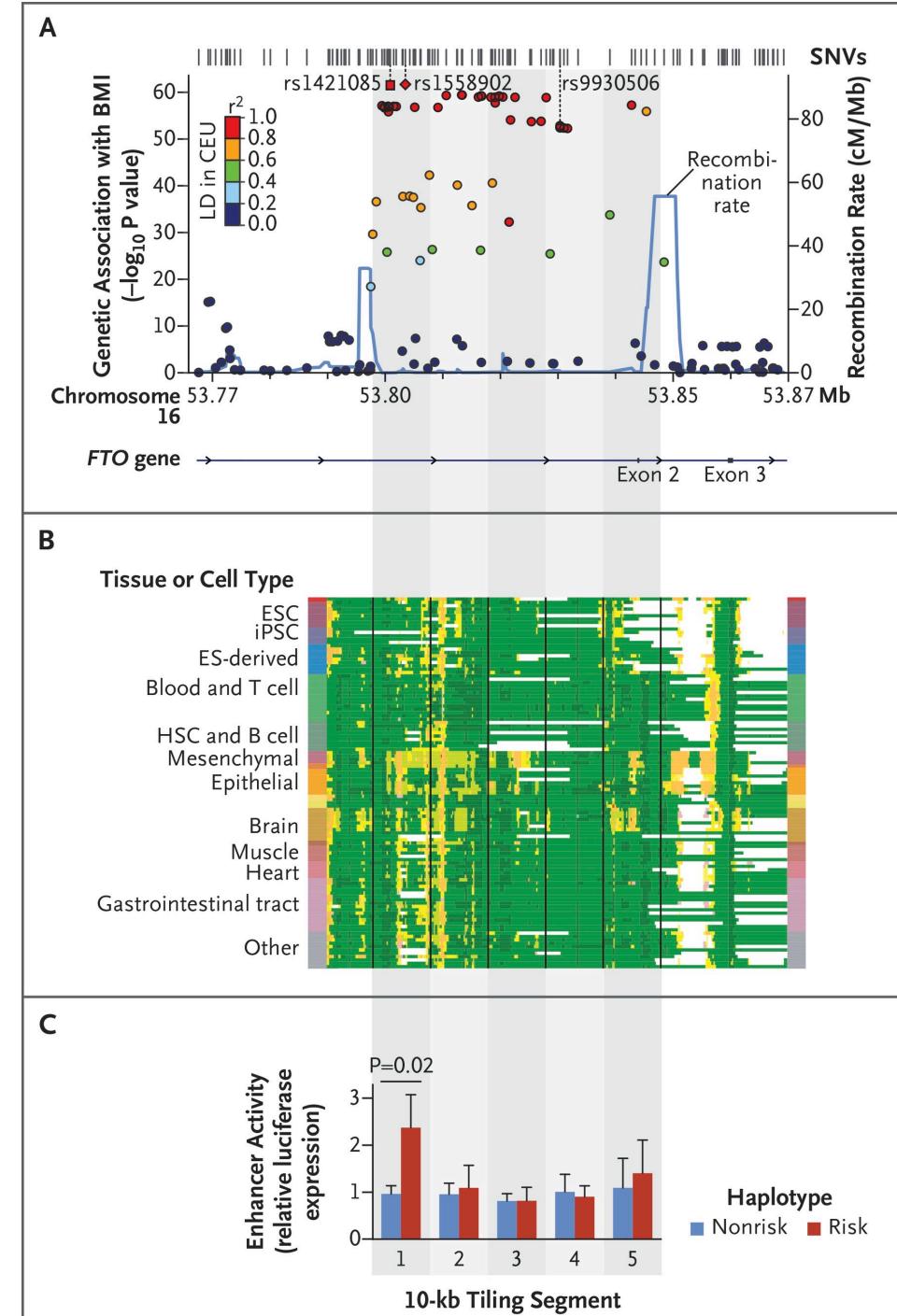


The case of *FTO*

“Fat mass and obesity-associated protein”

Risk associated variant (rs1421085) is in an enhancer region for *IRX3* and *IRX5*.

Claussnitzer et al. NEJM (2015)



Colocalization

Two big families of approaches

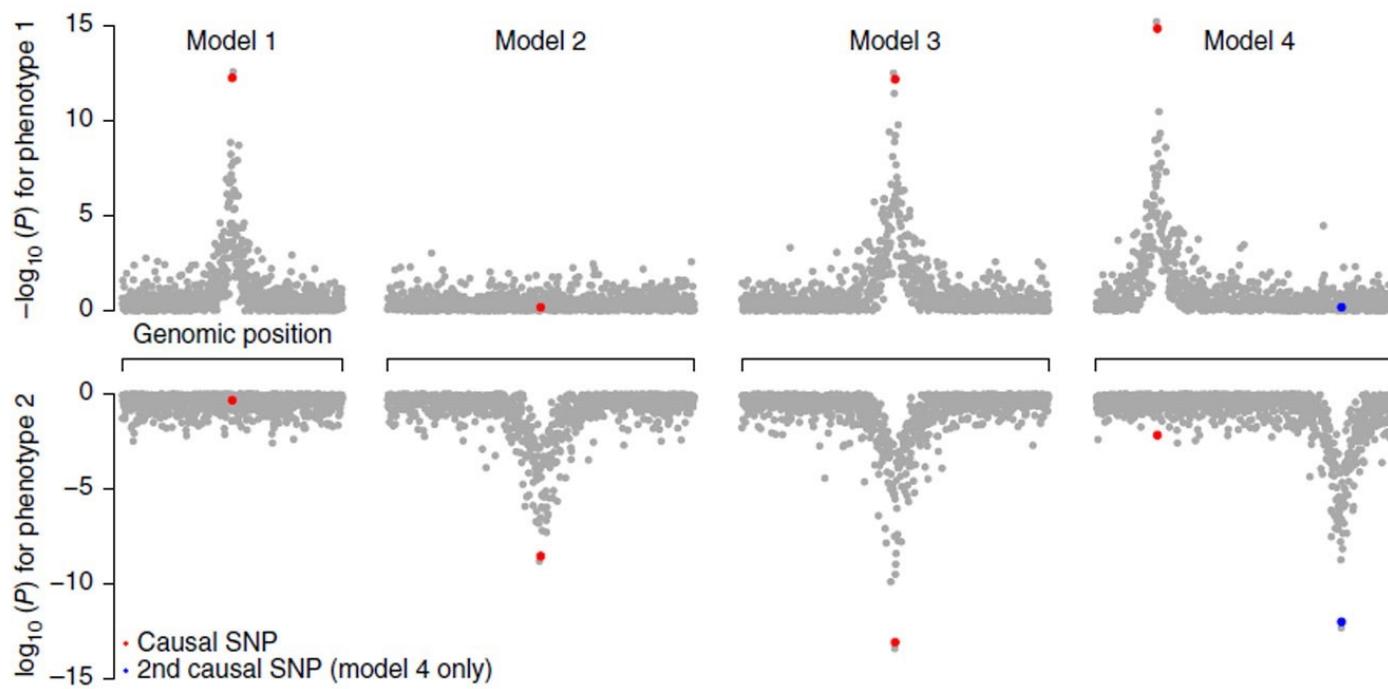
- 1) Enumeration colocalization: coloc method;
Giambartolomei et al. Plos Genet. (2014)
- 2) Proportional colocalization: HEIDI test (Zhu et al. - SMR) or
eCAVIAR (joint fine-mapping)

Coloc overview

Five hypotheses to test at given genomic locus

- **H0:** No association with either trait
- **H1:** Association only with trait 1
- **H2:** Association only with trait 2
- **H3:** Association with both traits but at separate causal variants
- **H4:** Association with both traits at a shared causal variant

H3 and H4 are the most interesting ones. H4 = colocalization!



nature genetics

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature genetics](#) > [analyses](#) > [article](#)

Published: 16 May 2016

Detection and interpretation of shared genetic influences on 42 human traits

[Joseph K Pickrell](#)✉, [Tomaz Berisa](#), [Jimmy Z Liu](#), [Laure Ségurel](#), [Joyce Y Tung](#) & [David A Hinds](#)

Nature Genetics **48**, 709–717 (2016) | [Cite this article](#)

SNP to Gene mapping - Summary

- Positional mapping + gene-based tests
Caveat: relevant genes may not be the closest ones (e.g., FTO vs IRX3/5)
- Adding omics-QTL data (MR or colocalization)
Caveat: relevant tissue/cell-type or omic may not be available

Part 3: Gene sets / Pathway enrichment

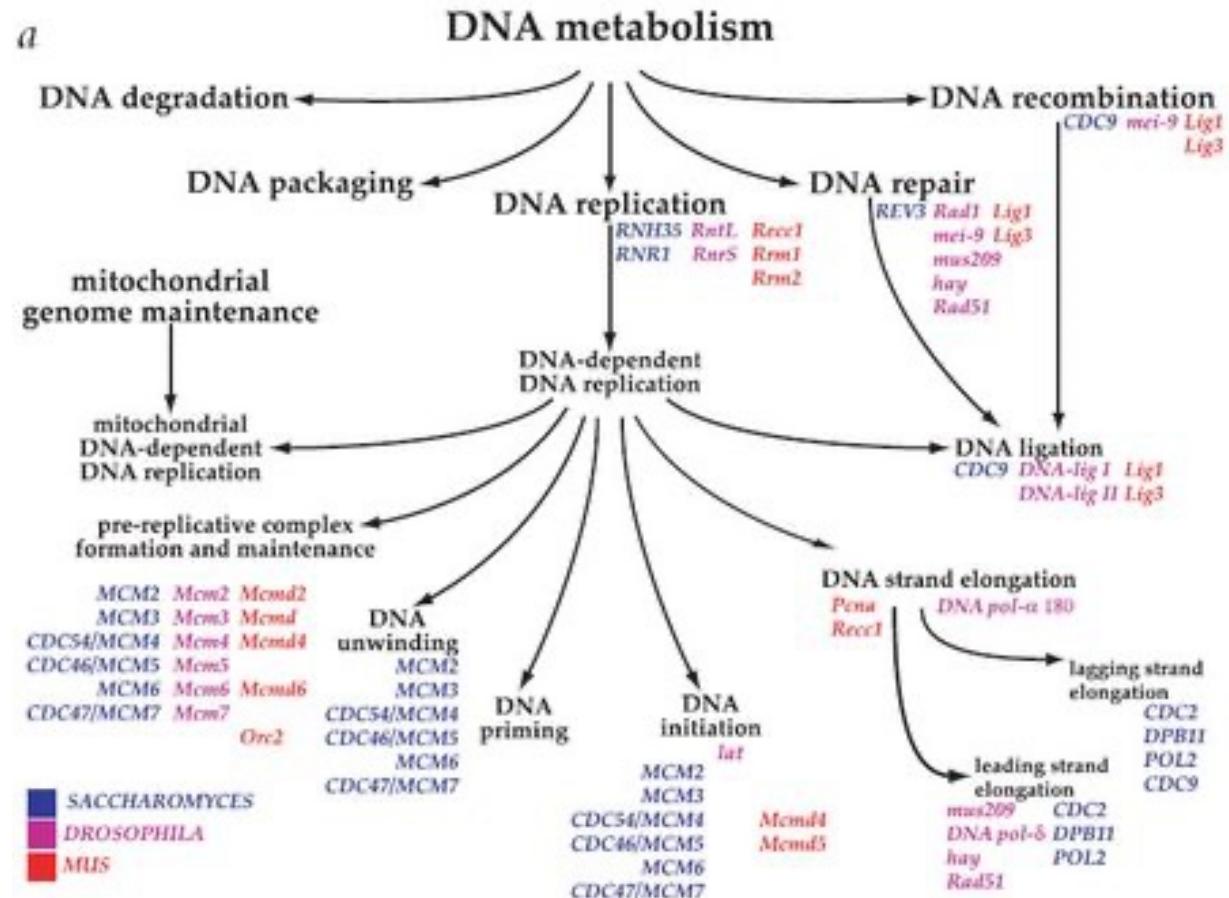
Credit: Slide content inspired from Timothy Thornton and Michael Wu
2017 SISG Lecture 4.

Gene sets / Pathway enrichment

- Most biological phenomena occur through the concerted expression of multiple genes (signalling pathways or functional relationships)
- Numerous databases organizing genes into groups exist (e.g., KEGG or Gene Ontology (GO) Consortium)

Gene Ontology Taxonomy

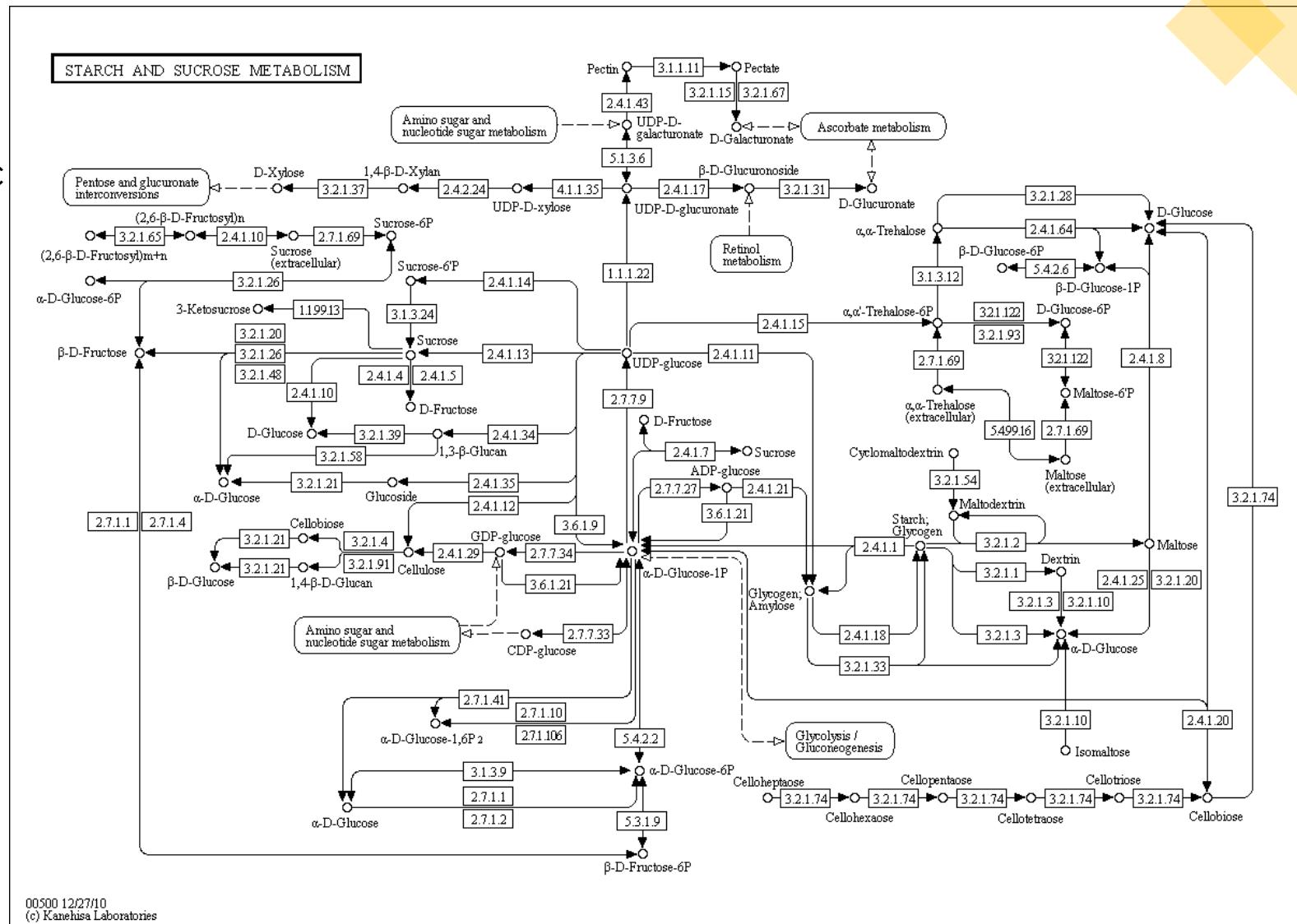
- Three principal ontologies: Biological Processes, Cellular Components, and Molecular Functional
- The graph has a hierarchy of terms (GO terms) from very broad (metabolism) down to more narrow levels (GTP biosynthesis)
- Each ontology and GO term has a comprehensive list of genes previously demonstrated to be associated with that ontology or GO term.



SACCHAROMYCES
DROSOPHILA
MUS

KEGG

- “Collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals” – Wiki
 - KEGG Pathways is network of gene pathways
 - Cleaner set of pathways than GO, but much smaller: emphasis on metabolic pathways though there are also disease and other trait related pathways.



A few methods to test enrichment

- Over-representation Analysis (ORA)
- Gene Set Enrichment Analysis (GSEA)
- Min P-value, Collapsing, Combined p-value, etc.

Over-representation Analysis

	Significant	Not significant	
In Pathway	X	X	
Not in Pathway	X	X	
Total			

Length and LD Bias

Generate a p-value for representation by using a test for independence:

- Fisher's Exact Test
- χ^2 -test
- Hypergeometric Test
- Binomial proportions z-test

Longer Genes capture more SNP signal by definition
Solution: use length-matched random sets of genes

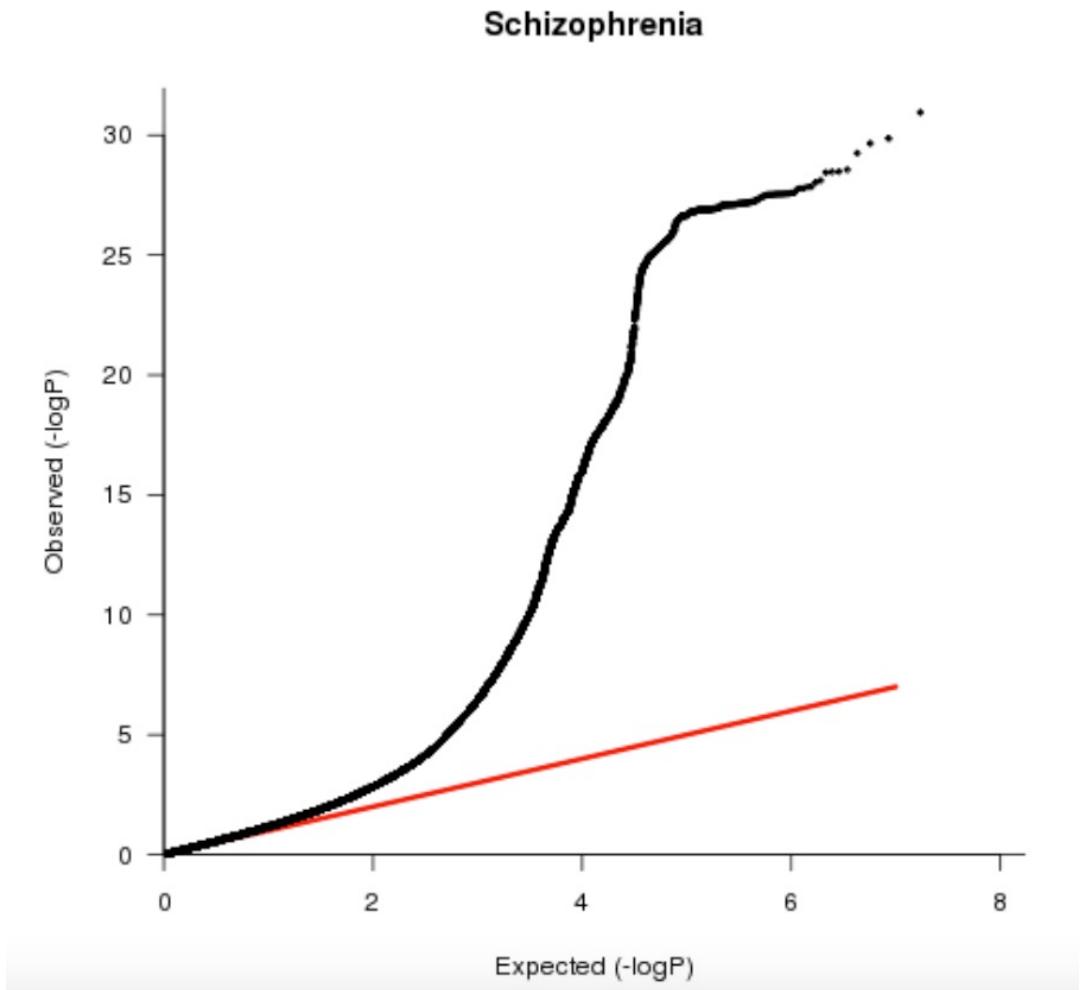
Another view...

Question: where in the genome is heritability concentrated?

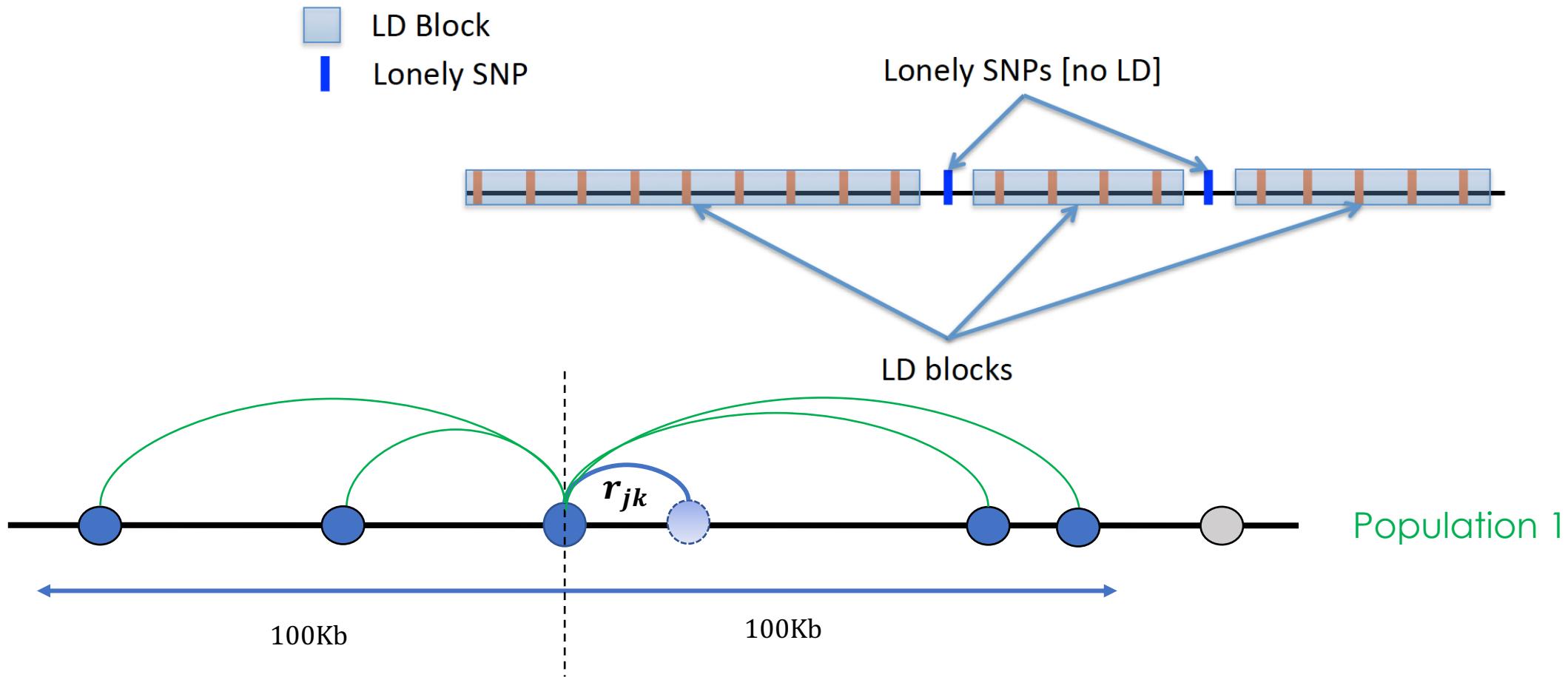
- Yang et al. (2012): coding vs non-coding regions?
- Gusev et al. (2011): regulatory elements (e.g., DHS, Histone marks)
- **Finucane et al. (2015) => GWAS summary statistics**

LD score regression

- Initial motivation: distinguish **polygenicity** from **confounding** (e.g., due to population stratification)
- Extension(s)
 - Estimation of SNP-based heritability and genetic correlations (Bulik-Sullivan 2014, 2015)
 - Functional Enrichment (Finucane 2015)
 - Estimation of polygenicity
 - Etc.



LD scores

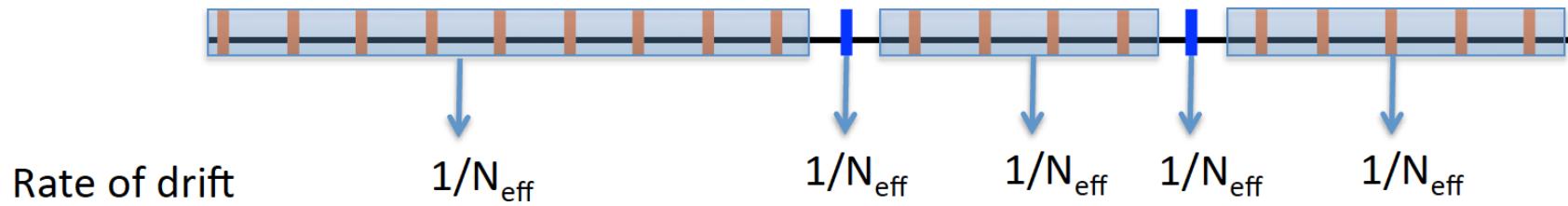


$$\text{LD score of SNP } j: \ell_j = \sum_{k=1}^M r_{jk}^2$$

Credit to Bullik-Sullivan (online lecture)

Under genetic drift...

- LD Block
- Lonely SNP

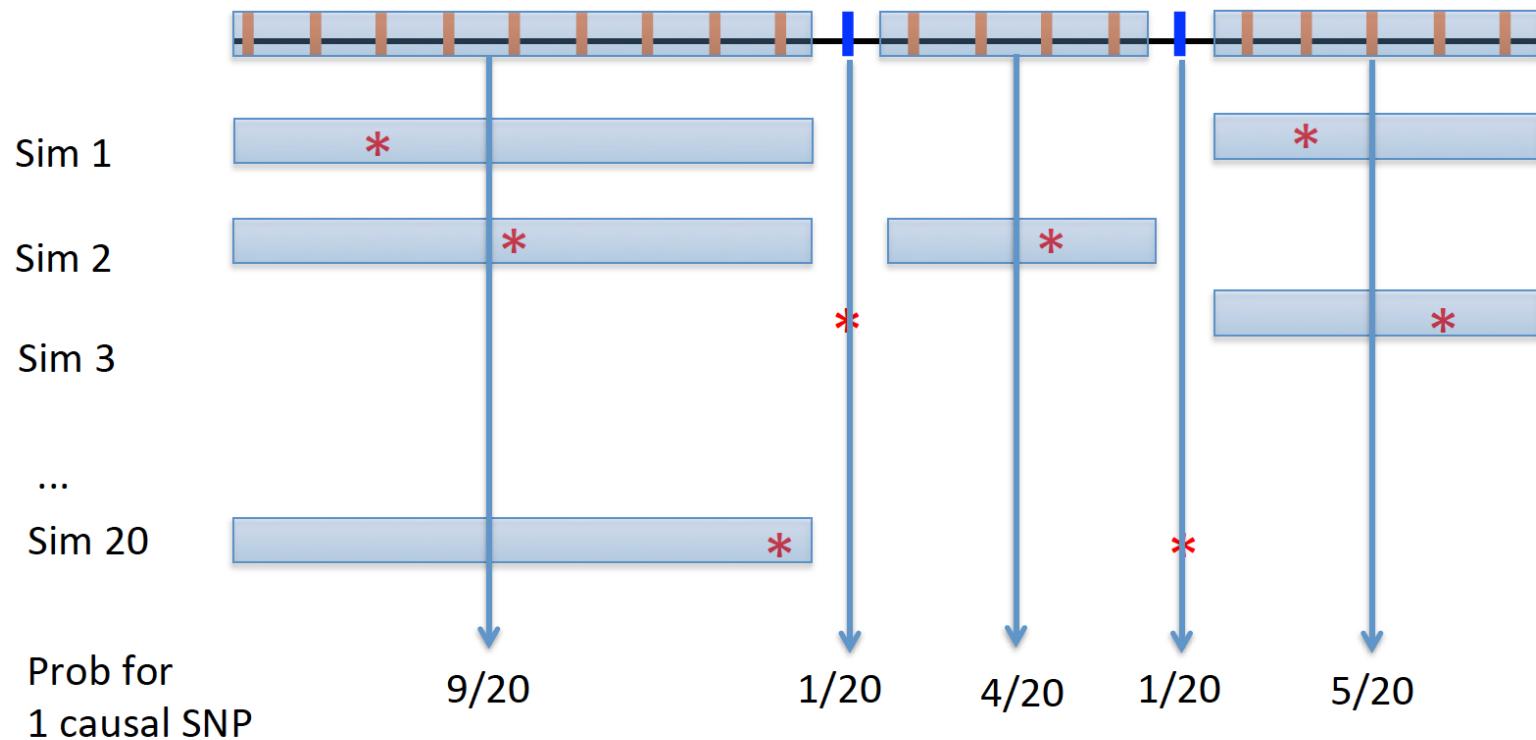


Under pure drift, LD is uncorrelated to magnitude of allele frequency differences between populations

...the more you tag, the more likely you are to tag a *causal variant!*

- LD Block
- Lonely SNP
- * Causal SNP

All SNPs in LD blocks w/ causal SNP have high chi-square

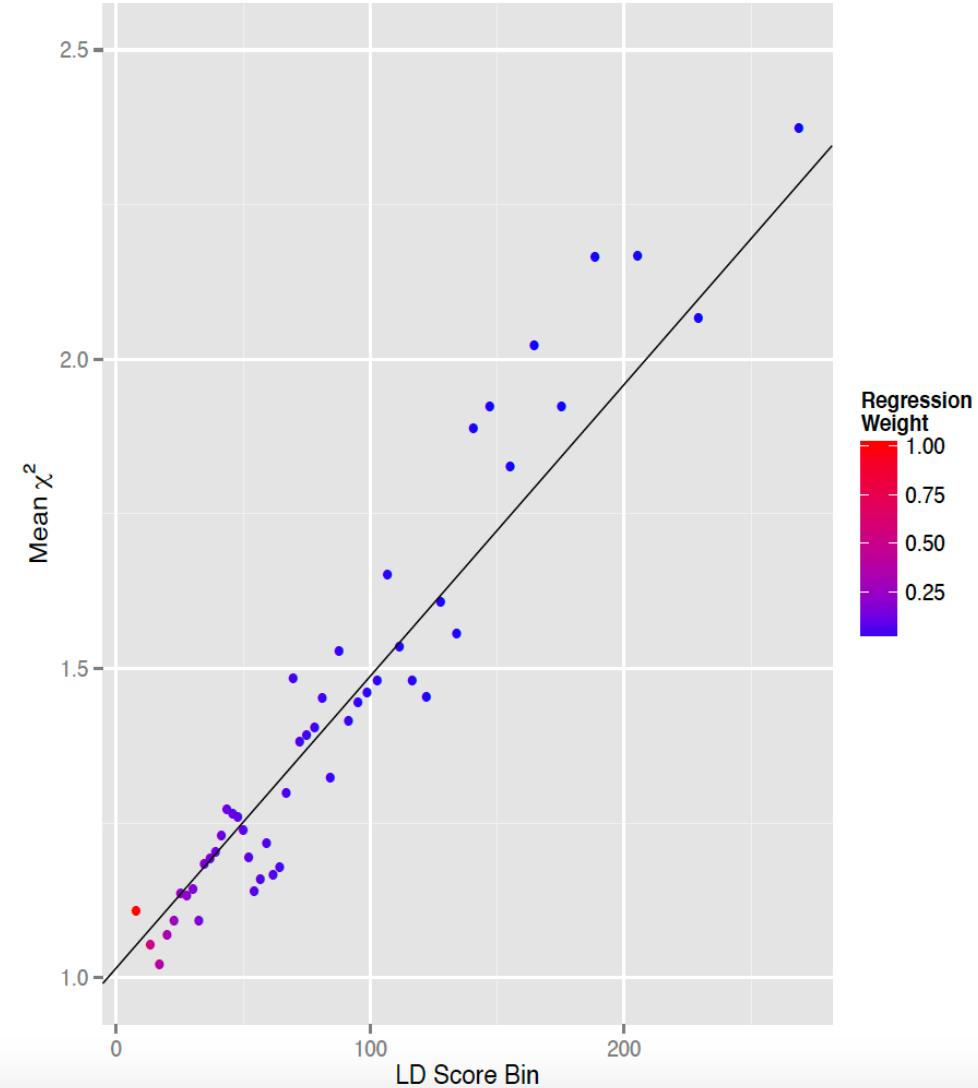
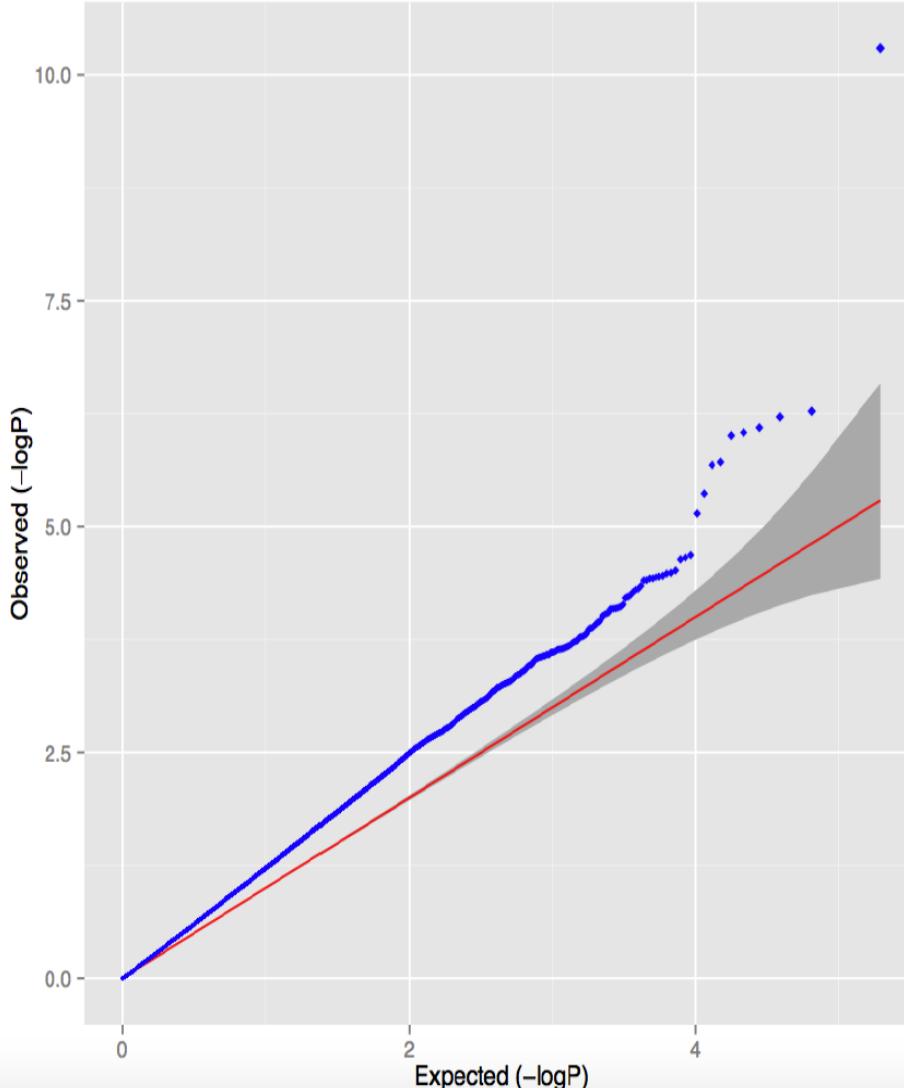


Key assumption
Each SNP explains the same amount of trait variance

Simulated Polygenicity

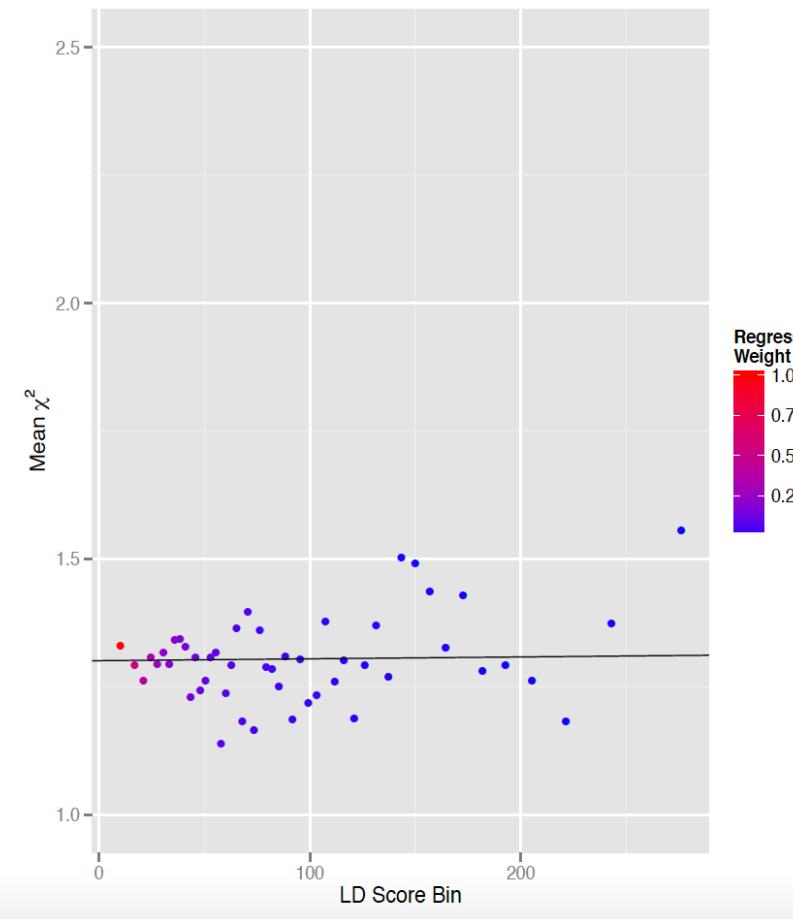
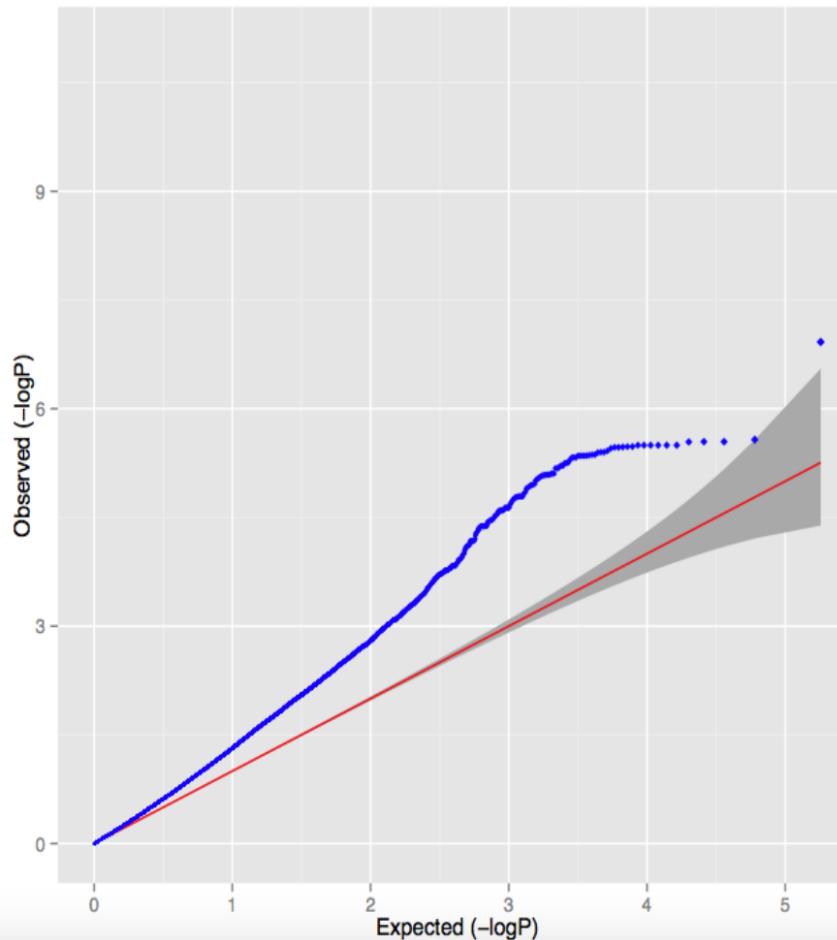
Credit to Bullik-Sullivan (online lecture)

- $\lambda_{GC} = 1.30$; LD Score Regression intercept = 1.02



Simulated population stratification (UK vs Sweden)

- $\lambda_{GC} = 1.30$; LD Score Regression intercept = 1.32



LD score regression theory

$$E[\chi_j^2 | \ell_j] = 1 + Na + \frac{Nh^2}{M} \ell_j$$

N is the GWAS sample size

$\frac{h^2}{M}$ is the average heritability explained per SNP.

$I = 1 + Na$ is the LD score regression intercept. Deviations from 1 indicate **confounding**.

Extension: Partitioned Heritability

<https://www.youtube.com/watch?v=-R130cCKd2A&t=10s>

Under the assumption that heritability varies linearly as function of certain functional annotations

$$E[\chi_j^2 | \ell_{jA_1}, \dots, \ell_{jA_K}] = 1 + Na + N \sum_{k=1}^K \tau_k \ell_{jA_k}$$

τ_k is the effect of annotation A_k and

$$\ell_{jA_k} = \sum w_k r_{jk}^2$$

is a weighted LD score between SNP j and SNPs within annotation k .

w_k can be 0/1 (binary: coding vs non-coding) or have continuous values (recombination rate).

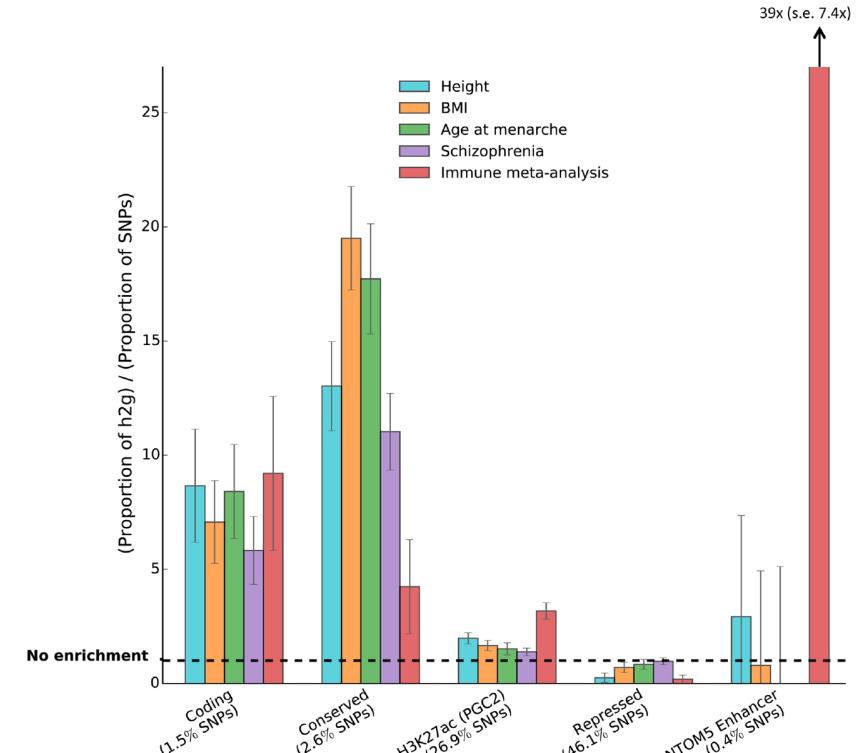
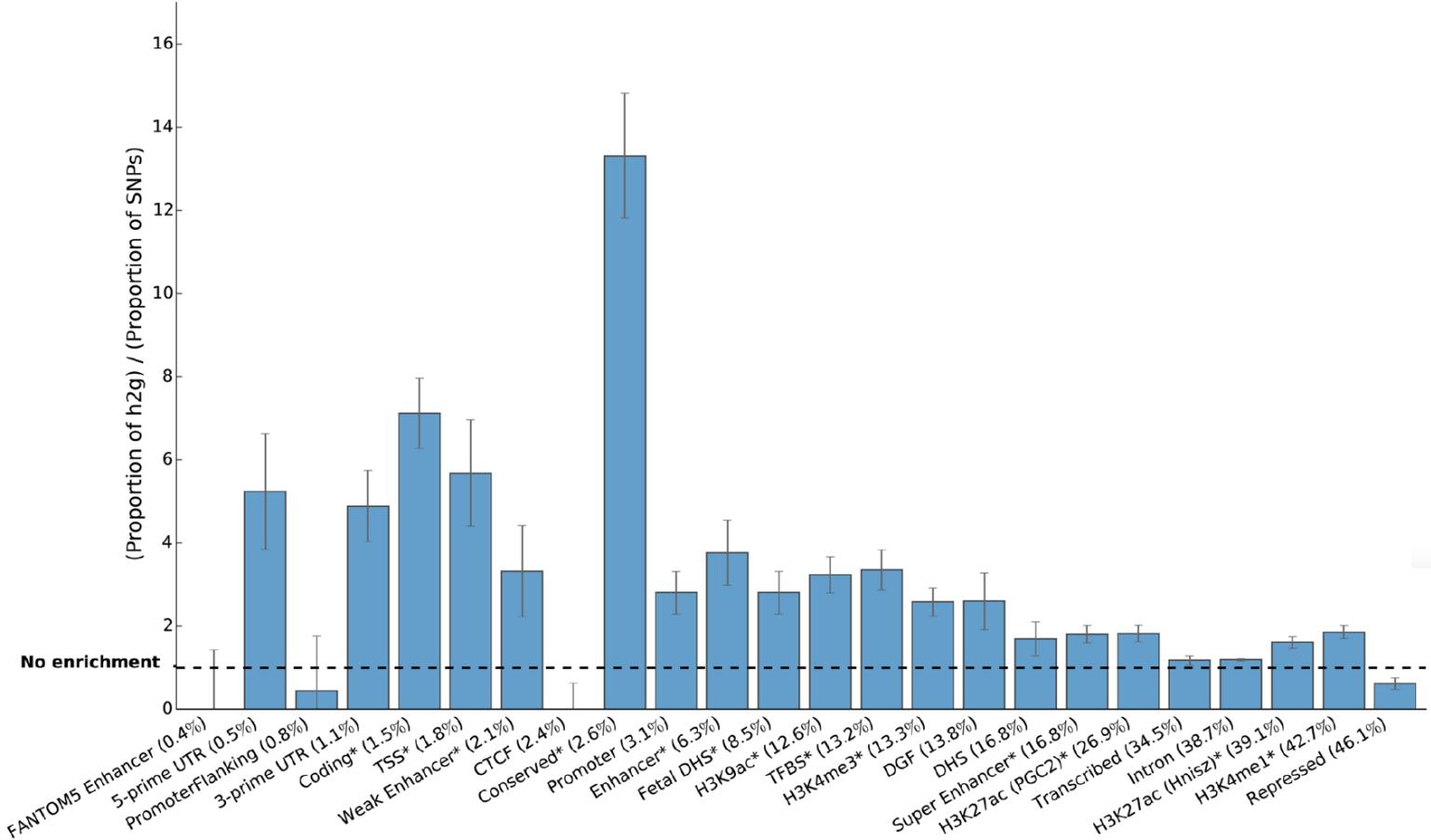
Applications

This framework can be used to prioritize annotations

- Tissues or cell types of interest (e.g., brain)
- Relevant functional class
- Pathways or gene sets

Example from Finucane (2015)

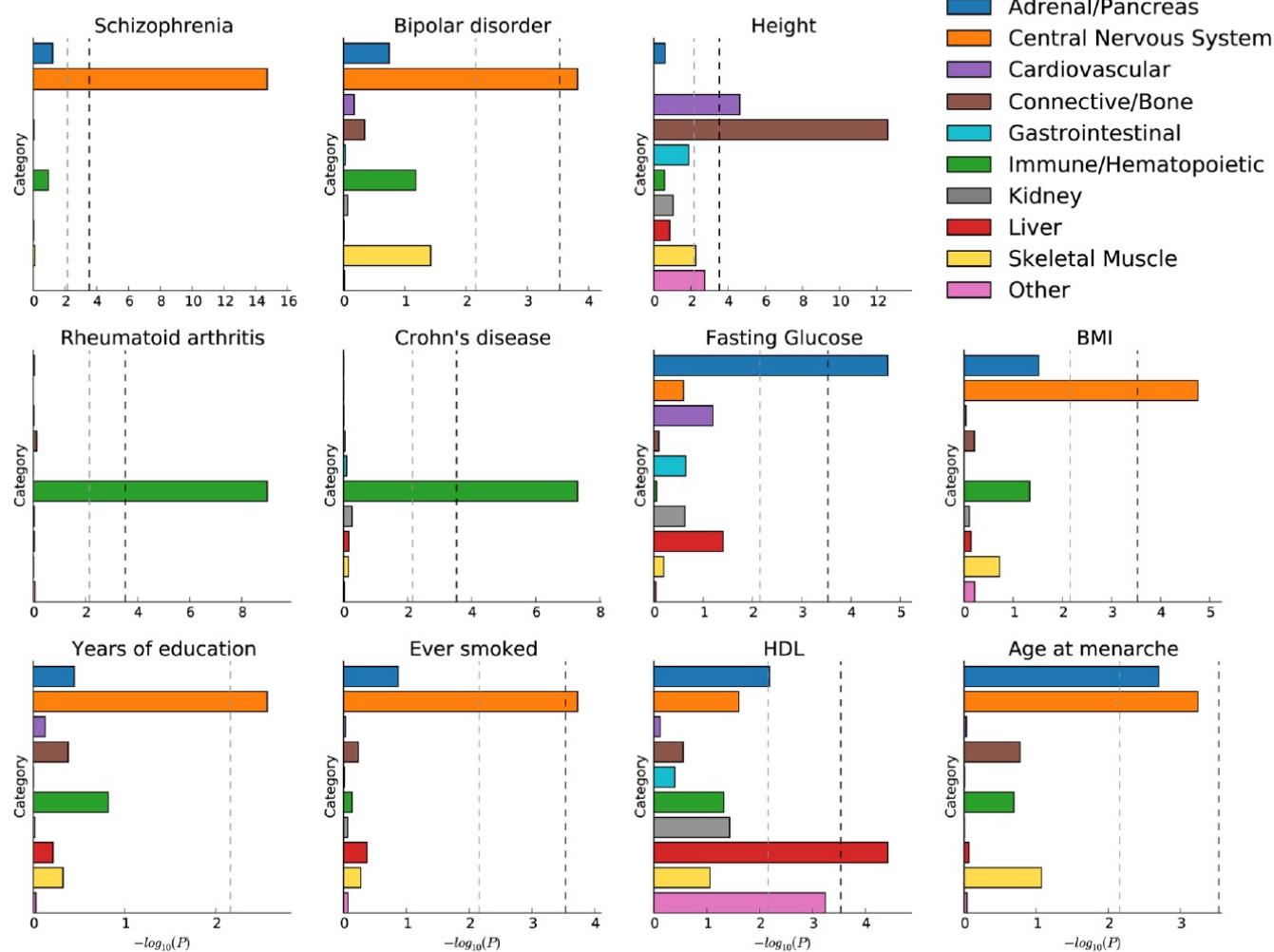
Enrichment across traits



LDSC can convert annotation effects (τ_k) into measures of enrichment

Example from Finucane (2015)

Tissue prioritization



Summary and conclusions

- LD score regression can be used to partition trait heritability across annotations
- Hybrid approach that uses both positional mapping and LD between focal SNPs and those in annotation



- The Polygenic Priority Score (PoPS) method is a proper extension (+other features)

Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases

Elle M. Weeks, Jacob C. Ulirsch, Nathan Y. Cheng, Brian L. Trippe, Rebecca S. Fine, Jenkai Miao, Tejal A. Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P. Fulco, Katherine C. Tashman, Francois Aguet, Taibo Li, Jose Ordovas-Montanes, Christopher S. Smillie, Moshe Biton, Alex K. Shalek, Ashwin N. Ananthakrishnan, Ramnik J. Xavier, Aviv Regev, Rajat M. Gupta, Kasper Lage, Kristin G. Ardlie, Joel N. Hirschhorn, Eric S. Lander, Jesse M. Engreitz, Hilary K. Finucane

doi: <https://doi.org/10.1101/2020.09.08.20190561>