THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Analyses Downstream of GWAS

2nd Australian UK Biobank Research Symposium, February 9-13 2026

UK Biobank Instructional Workshop

Alesha Hatton

Slide credits
UQ Genetics and Genomics Winter School
International Statistical Genetics Workshop
Summer Institute in Statistical Genetics

Moving beyond association mapping

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

Issues:

- SNPs are correlated (LD)
- Polygenic traits influence by many variants/genes
- GWAS effects small magnitude (need large sample sizes)
- Most GWAS hits are outside genes or in non-coding regions of genome, with unclear mechanisms

What to do after your GWAS?

- Make it bigger! **Meta analysis**
- How much trait variance is captured by genetics (how much is confounding?) **SNP heritability**
- How many distinct associations there are **Conditional analysis**
- **Fine-map** your results (getting as close as possible to where “causal” variants are)
- Find what are the relevant genes? **Gene based tests**
- Find what are the relevant biological pathways? **Gene-set/Pathway enrichment**
- Do these variants influence multiple traits? **PheWAS and UKB phenotypes**

…..

GWAS Meta-analysis

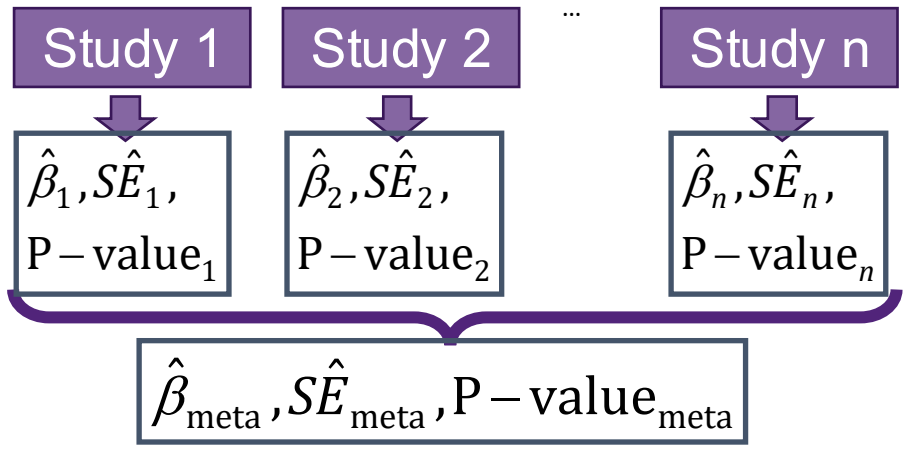
Combine data across studies

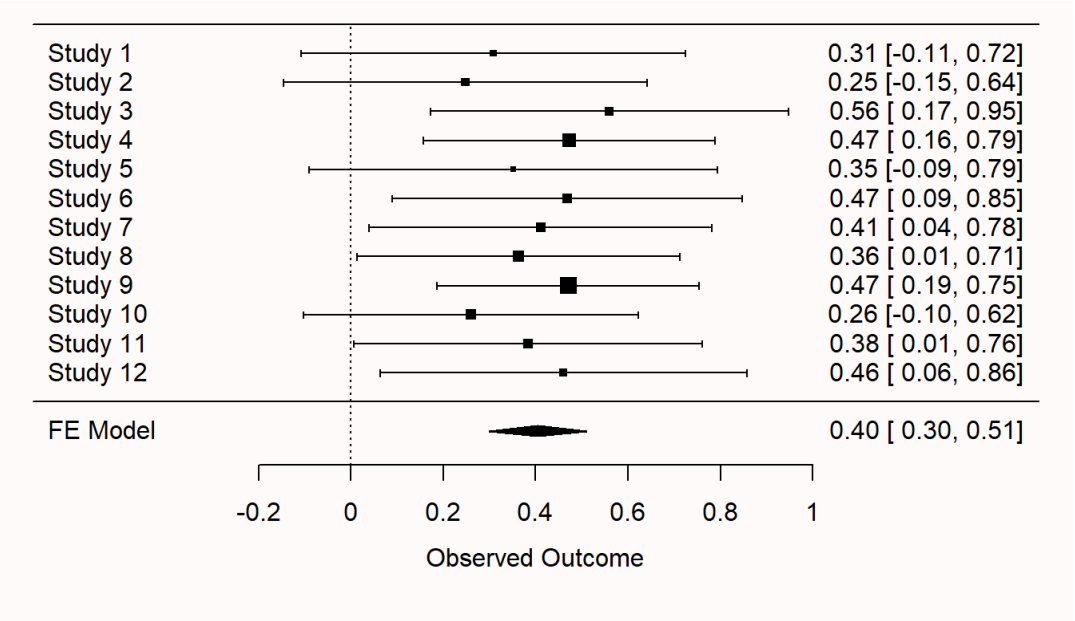
Aims:

- Increase sample size (power) without sharing individual-level data
- Explore differences between studies

Considerations

- Unified analysis protocol
- **Quality control of summary statistics**
- Trait definition
- Heterogeneity
- Independence of the samples



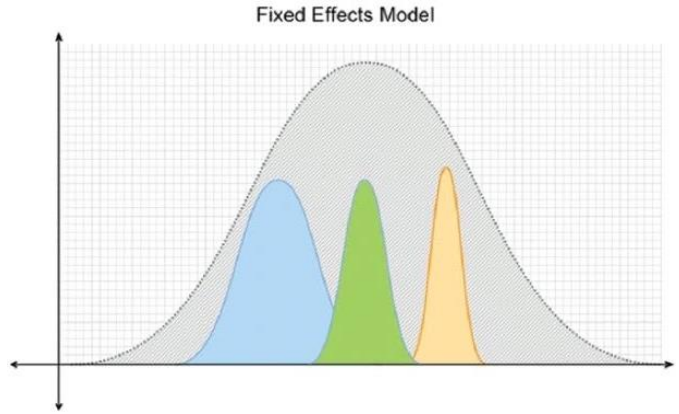


Fixed effect vs Random effects model

Fixed-effect model (more common for GWAS)

- Assume SNP has a single “true” effect on the trait across all cohorts
- Differences between studies are due to sampling error only

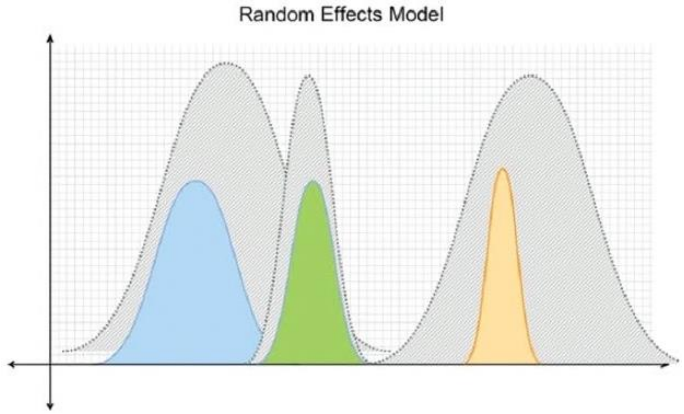
The ‘combined effect’ is the estimate of the true effect



Random-effects model

- Assume the true effect for a SNP varies between cohorts
- Studies are assumed to be a random sample reflecting the distribution of true effects
- Error in our estimate is due to within study sampling error and between study variance

The ‘combined effect’ is the mean of the normal distribution



- Fixed effects is more powerful than random effects in general but also more sensitive to errors in trait scaling and phenotype heterogeneity

Inverse Variance Weighted Fixed Effect model

- Estimates from each study are weighted by the inverse of the variance of the effect estimate ($1/SE^2$)
- Larger studies (with smaller SEs) are given more weight

$$\beta = \frac{\sum_i \beta_i w_i}{\sum_i w_i}$$

β_i: effect estimate for study i

w_i: weight for study i, given as $\frac{1}{se_i^2}$

se_i: standard error for study i

- Software – METAL

Other types of meta-analysis

Modelling multi-ancestry heterogeneity

- MR-MEGA

"*Modeling allelic effects as a function of axes of genetic variation, derived from a matrix of mean pairwise allele frequency differences between GWAS*"

JOURNAL ARTICLE

Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution

Reedik Mägi, Momoko Horikoshi, Tamar Sofer, Anubha Mahajan, Hidetoshi Kitajima, Nora Franceschini, Mark I. McCarthy, COGENT-Kidney Consortium, T2D-GENES Consortium, Andrew P. Morris

Human Molecular Genetics, Volume 26, Issue 18, 15 September 2017, Pages 3639–3650,

Modelling multiple phenotypes

- MTAG
- GenomicSEM

Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

Andrew D. Grotzinger, Mijke Rhemtulla, Ronald de Vlaming, Stuart J. Ritchie, Travis T. Mallard, W. David Hill, Hill F. Ip, Riccardo E. Marioni, Andrew M. McIntosh, Ian J. Deary, Philipp D. Koellinger, K. Paige Harden, Michel G. Nivard & Elliot M. Tucker-Drob

Nature Human Behaviour **3**, 513–525 (2019) | Cite this article

Multi-trait analysis of genome-wide association summary statistics using MTAG

Patrick Turley, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A. Furlotte, 23andMe Research Team, Social Science Genetic Association Consortium, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, David Laibson, David Cesarini, Benjamin M. Neale & Daniel J. Benjamin

Nature Genetics **50**, 229–237 (2018) | Cite this article

What proportion of trait variation is explained by genetics?

Is the signal due to polygenicity or confounding (e.g. population stratification)?

Answer both of these questions with summary statistics using LDSC

- Requires us to account for
  - (1) LD
  - (2) confounding (not all signal is due to true effects)

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

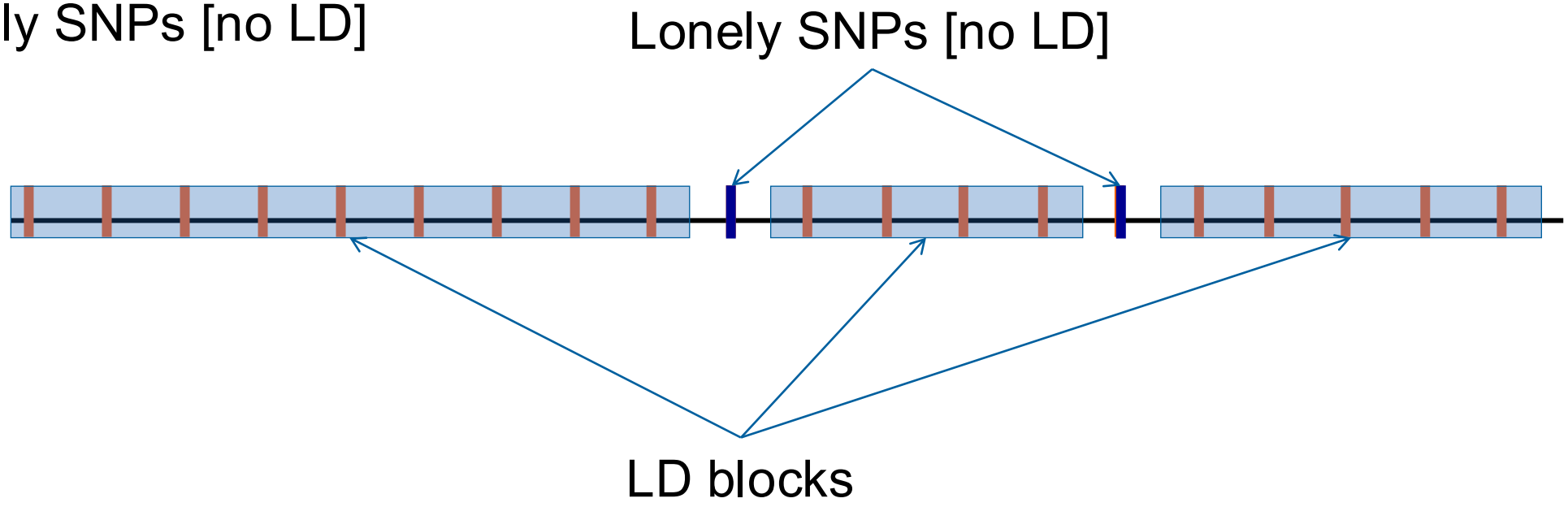
Nature Genetics **47**, 291–295 (2015) | Cite this article

How does LD shape association?

How does LD shape association?

LD blocks

Lonely SNPs [no LD]



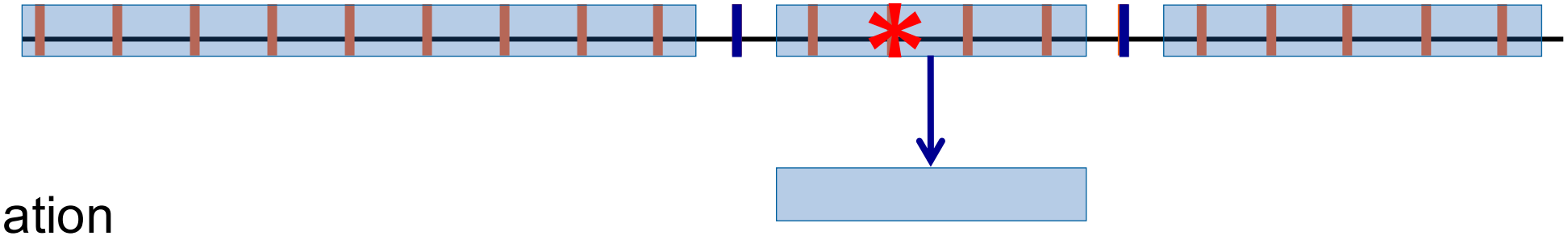
LD score of SNP j : $l_j = \sum_{k \neq j} r_{jk}^2$

How does LD shape association?

LD blocks

Lonely SNPs [no LD]

* Causal variants



Association

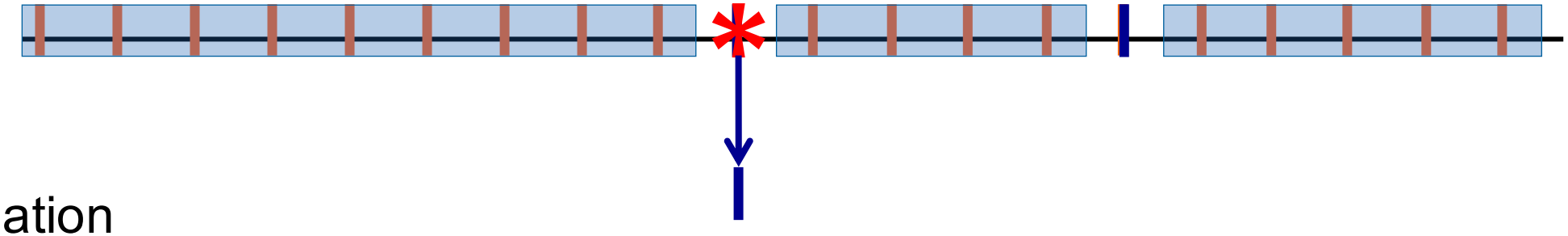
All markers correlated with a causal variant show association

How does LD shape association?

LD blocks

Lonely SNPs [no LD]

* Causal variants



Association

Lonely SNPs only show association if they are causal

How does LD shape association?

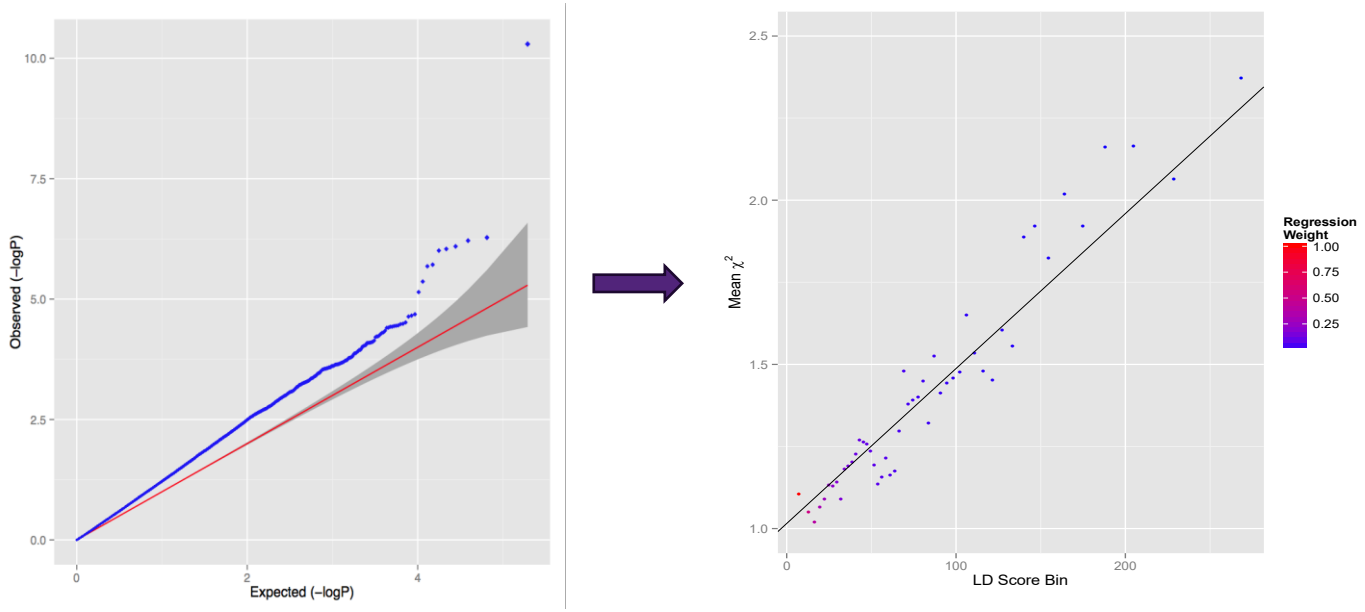
LD blocks

Lonely SNPs [no LD]

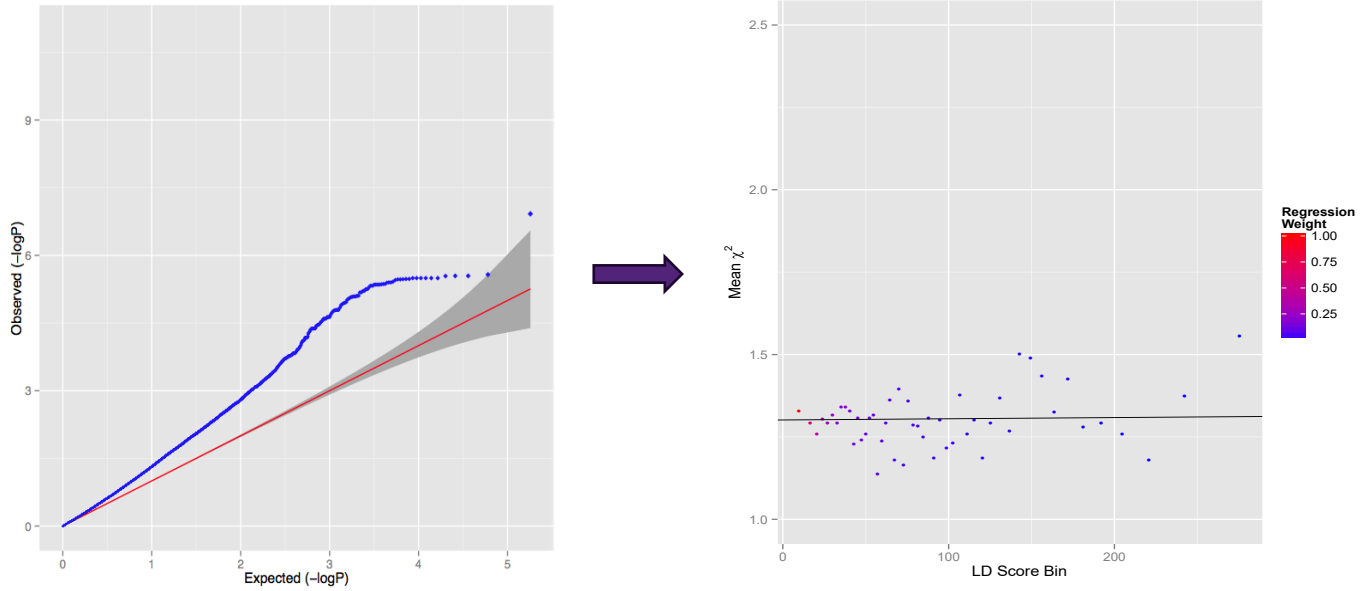
* Causal variants

The more you tag, the more likely you are to tag a causal variant

Key assumption *each SNP explains the same amount of trait variance*



Simulated polygenicity



Population stratification

LD score regression theory

$$E[\chi_j^2] = 1 + Na + \frac{h_g^2 N}{M} l_j$$

N is the GWAS sample size

$\mathbf{1 + Na}$ is the LD score regression intercept. Deviations from 1 indicate confounding

$\frac{\mathbf{h_g^2}}{\mathbf{M}}$ is the slope of the regression - average heritability explained per SNP

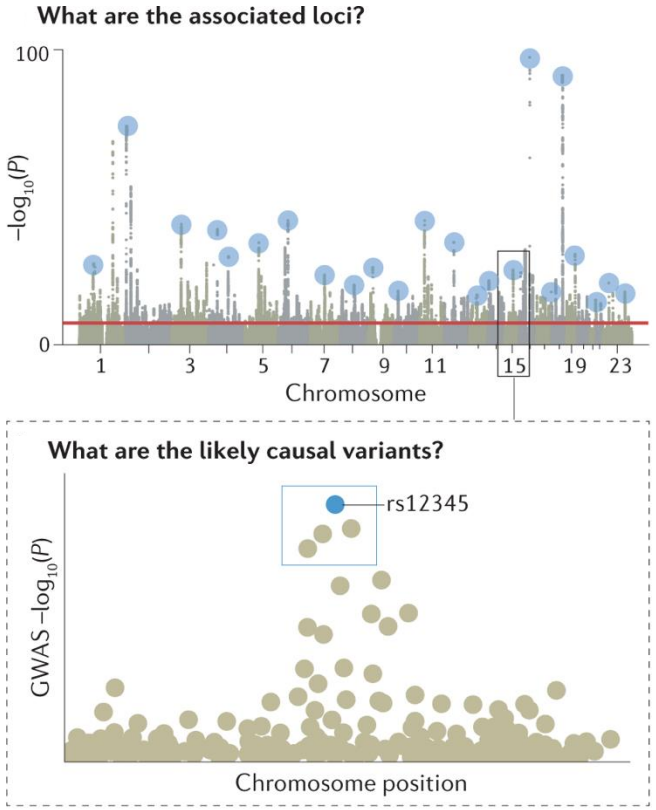
Caveats and extensions

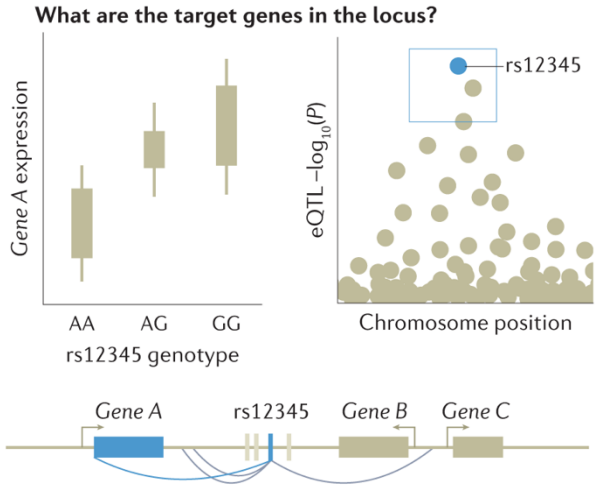
- Wont capture genetic variation that is not tagged well by common variation
- Matched LD reference

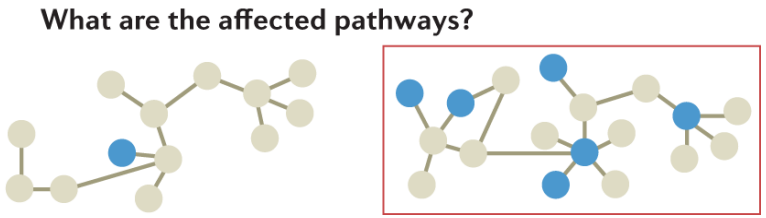
Extensions

- *Estimation of genetic correlations (Bulik-Sullivan 2014, 2015)*
- *Functional Enrichment (Finucane 2015)*
  - *Tissues or cell types of interest (e.g., brain)*
  - *Relevant functional class*
  - *Pathways or gene sets*

SNP → Gene → Pathway







Uffelmann, E., *et al.* Genome-wide association studies. *Nat Rev Methods Primers* (2021)

Counting the number of distinct associations

Helps to understand the genetic architecture of a trait

Useful for downstream analysis (e.g. MR)

Classical algorithm is LD clumping (PLINK)

- *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*

  - Overlooks independent variants if they are in close proximality and in LD
  - The relationship between window size and LD is dependent on genetic ancestry

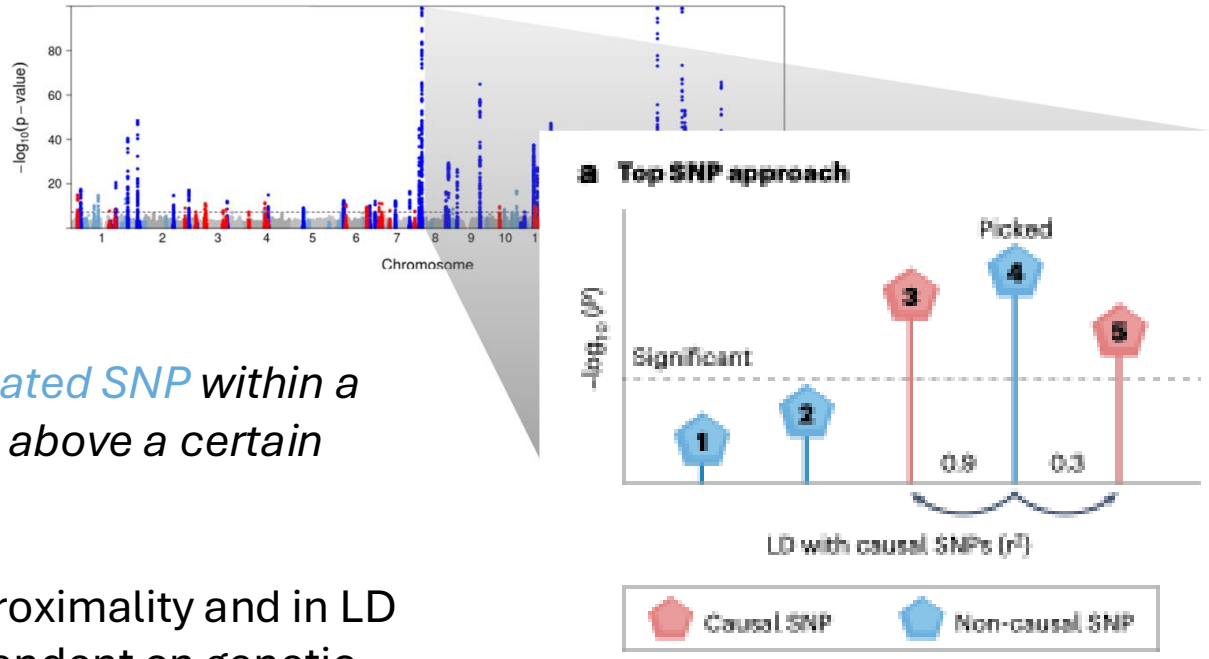
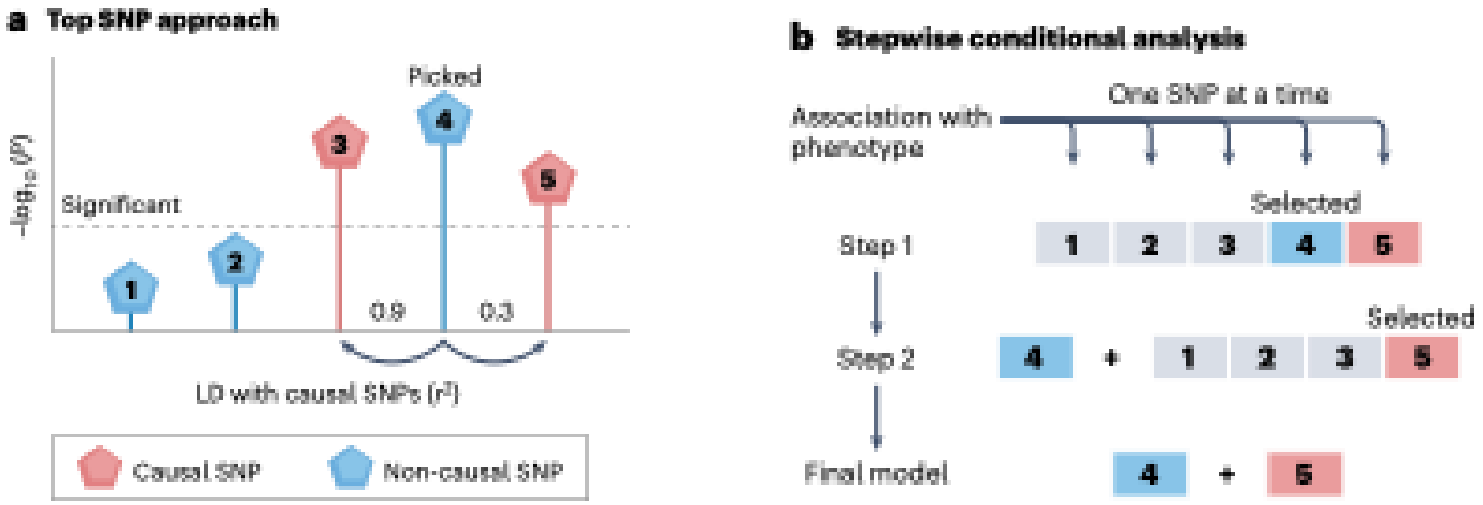


Figure from Li and Zhou 2025 Nature Reviews Genetics and Willer et al. Nat Genet 2013

(Approximate) Conditional and Joint Analysis (COJO)

(Yang et al 2012 Nature Genetics, implemented in GCTA software)

- Mimics stepwise variable selection
  1. First selects the top variant
  2. Iteratively tests for additional variants, while adjusting for those already included in the model
  3. Until no further variant reaches a pre-specified significance threshold



- Doesn't assess the statistical uncertainty of the identified variants because it does not consider other possible combinations of variants that could explain the association equally well, or even better

(Bayesian) Fine-mapping

- **Aim**: Identify minimal set of variants and ascribe probabilities of causality
- It is straightforward to make causality inference based on the posterior probability

Two key statistical outputs

- **PIP**: probability of the SNP being causal given the data and all other SNPs $PIP_j = Pr(b_j \neq 0 \mid X, Y)$
- **Credible set:** a set of SNPs that jointly capture a causal variant with a given probability

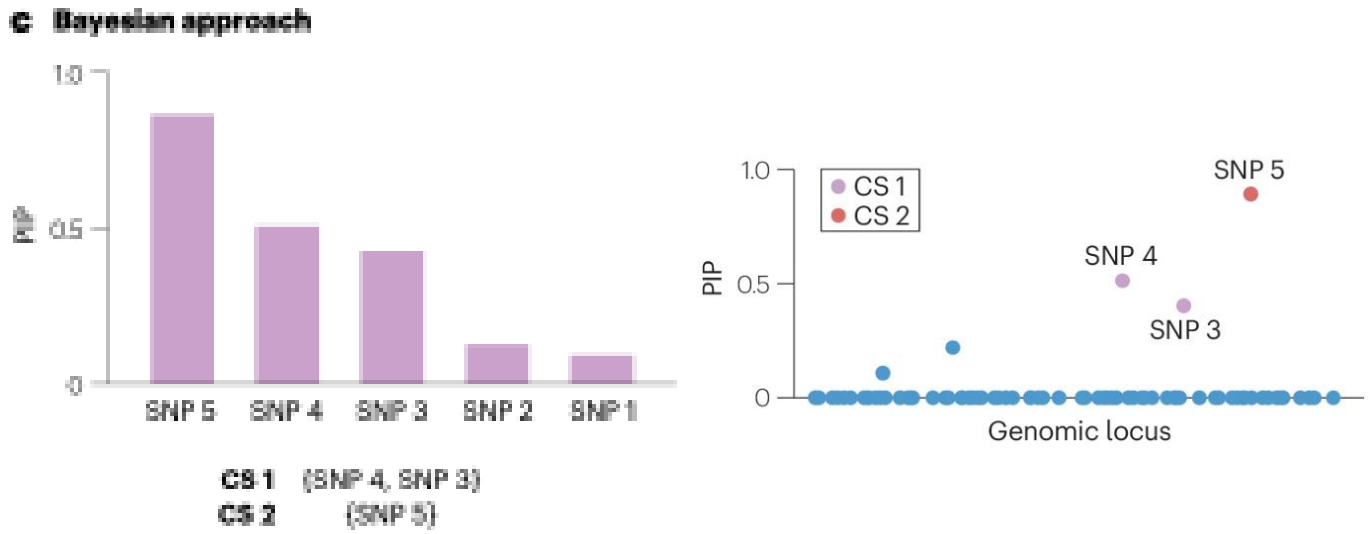
- No need multiple test adjustment

Posterior inclusion probability

SNP	PIP
SNP 1	0.1
SNP 2	0.2
SNP 3	0.4
SNP 4	0.5
SNP 5	0.9

Credible sets

CS	SNPs
CS 1	{SNP 4, SNP 3}
CS 2	{SNP 5}



Li, Zhou, Towards improved fine-mapping of candidate causal variants. *Nat Rev Genet* (2025)

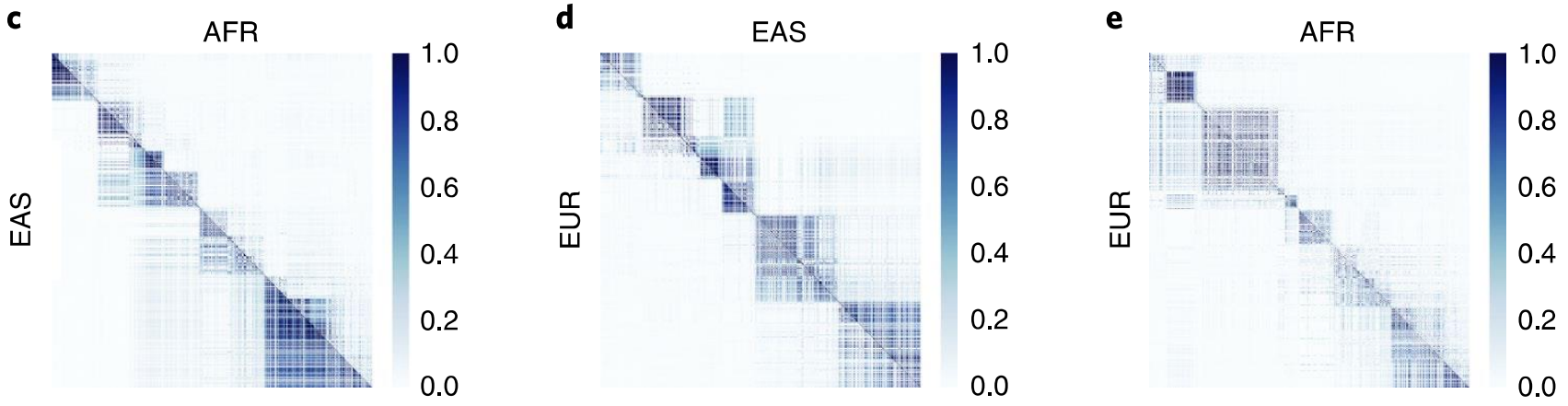
Fine-mapping methods

Differ in the prior assumption of distribution of true SNP effects and the algorithm for model fitting

- BIMBAM (Servin and Stephens, 2007)
- CAVIAR (Hormozdiari, 2014)
- PAINTOR (Kichaev, 2014)
- CAVIARBF (Chen, 2015)
- FINEMAP (Benner, 2016)
- SuSiE (Wang, 2020)
- Susie-inf & FINEMAP-inf (Cui, 2024)
- SBayesRC (Wu, 2025 medRxiv)

Challenges

- The causal variants have been measured
  - Not true for rare variants or structural variants!
- The LD reference samples are from the same population as the GWAS samples
  - No systematic differences in LD – same ancestry!
  - Minimum sampling variance in LD – LD reference sample size cannot be too small
- Meta-analysis fine-mapping is tricky, see Kanai et al. 2022 Cell Genomics.

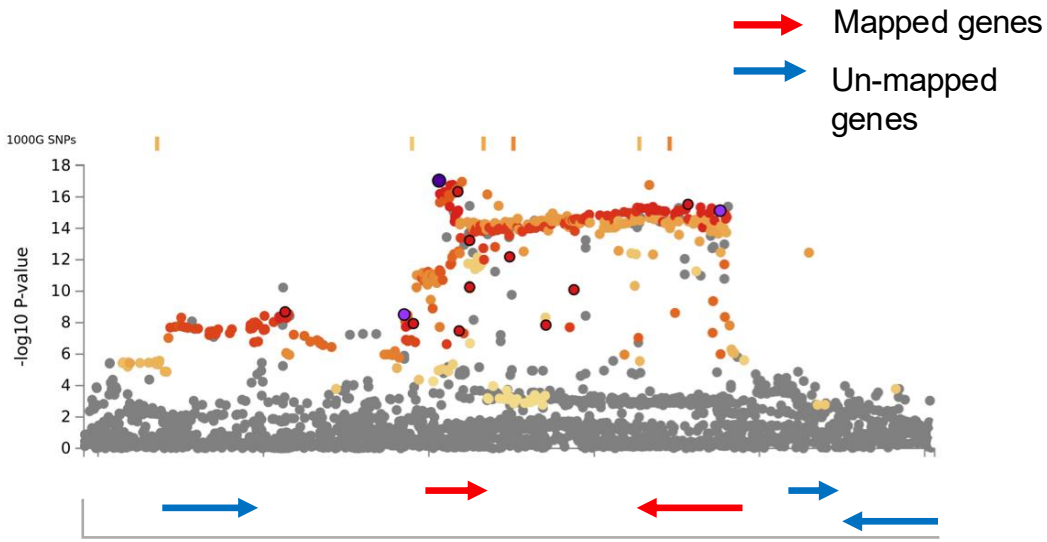


Martin et al. 2019 Nature Genetics.

Mapping SNPs to genes

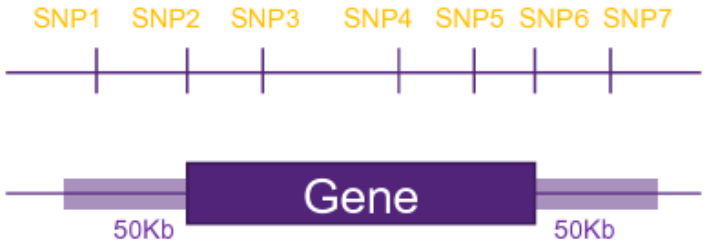
Identify functional impact of GWAS hits

- Positional mapping: *relevant genes must be nearby*
- Gene based test - joint association effect of all SNPs in a gene



Gene-based tests

Unit of analysis is the gene



Pros

- Reduced multiple-testing burden

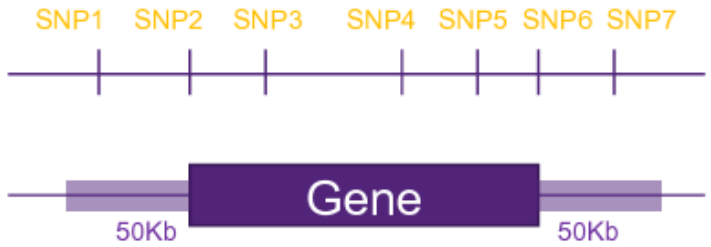
  (millions of SNPs vs ~22k genes)
- Detect effects consisting of multiple weaker associations

Cons

- Disregards regulatory (often non-genic) information when based on location-based annotation
- Not all methods account for masking effects

Gene based analysis

Summary level data and LD reference



- **Step 1:** allocate SNPs to genes (positional mapping e.g. within 100kb of TSS)

- **Step 2:** aggregate Z-scores

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

Several methods available

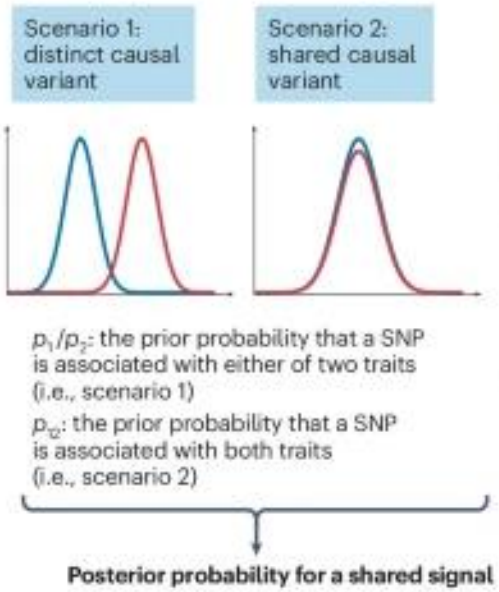
MAGMA, VEGAS (PLINK), mBAT-combo (GCTA)

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

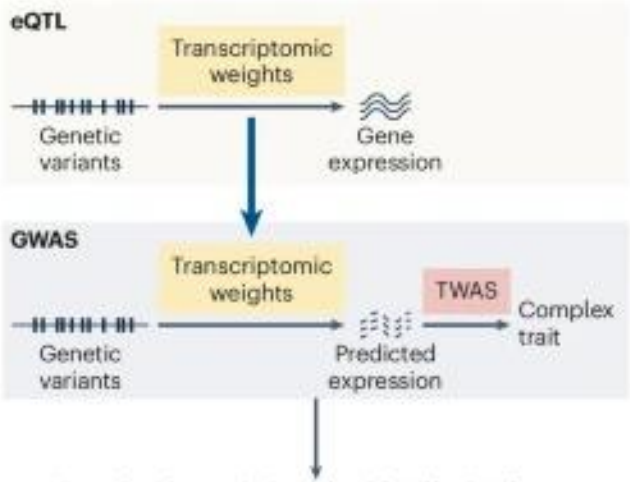
Mapping SNPs to genes using to molecular QTLs

Hypothesis: GWAS SNPs are eQTLs (or pQTLs) for their causal gene

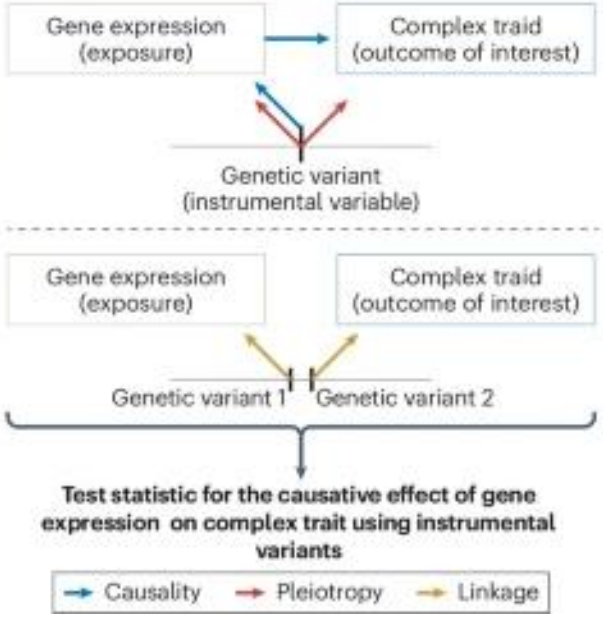
Colocalization



TWAS



MR



Gene implication methods often produce inconsistent results - increase robustness by using multiple approaches

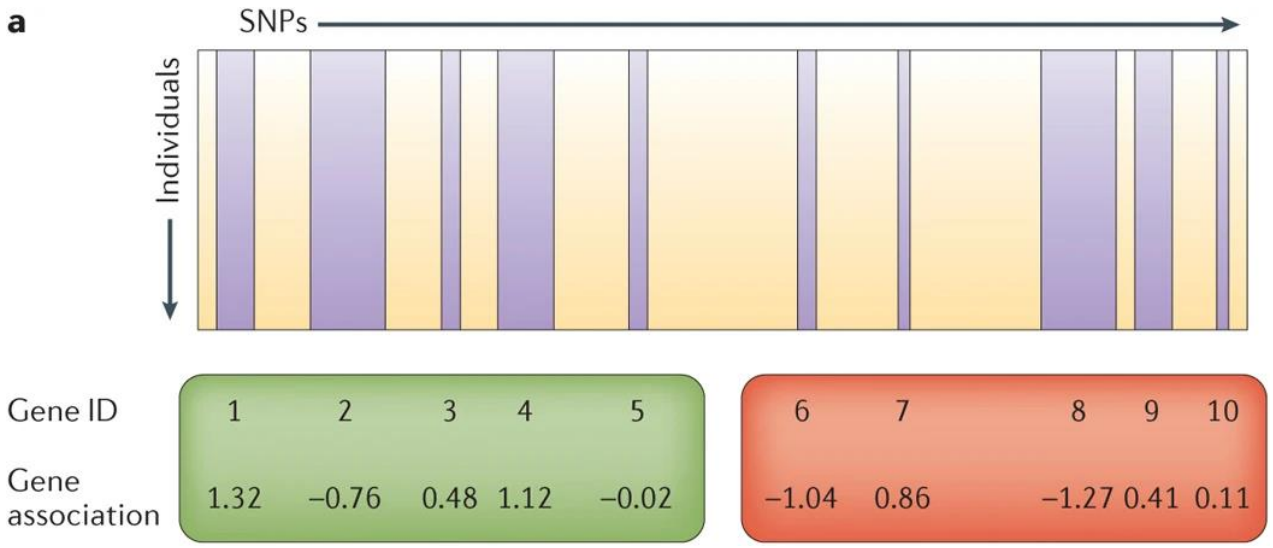
Liu, F., *et al.* Mitigating inconsistencies in GWAS follow-up analyses with LocusCompare2. *Nat Genet* (2025)

Gene-set analysis (pathway enrichment)

Most biological phenomena occur through the concerted expression of multiple genes

Gene set - any group of genes that share a particular property e.g. same pathway, same protein family

Gene set analysis - determine whether that property of the gene set has a role in the phenotype of interest.



Leeuw et al Nat Rev Gene 2016

Gene-set analysis

1. **Self-contained analysis:**
   - null hypothesis: none of the genes in the gene set are associated with phenotype.
   - Tests if genes in a gene-set are jointly associated with the phenotype of interest
   - Only considers genes in the gene set

Self-contained analysis:
$\theta_S > \emptyset?$

2. **Competitive analysis:**
   - tests if genes in a gene-set more strongly associated with the phenotype than other genes
   - Considers all genes in the data
   - Joint association of genes in the gene set is greater than the association of genes not in the gene set

Competitive analysis:
$\theta_S > \theta_0?$

Leeuw et al Nat Rev Gene 2016

Gene-set analysis

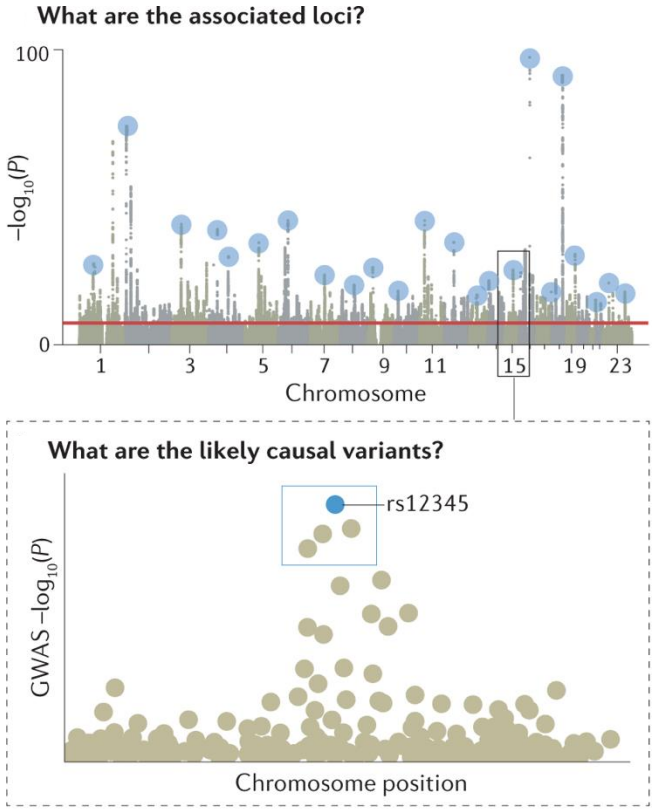
Numerous databases organizing genes into groups exist – intended to represent the current state of knowledge in biology

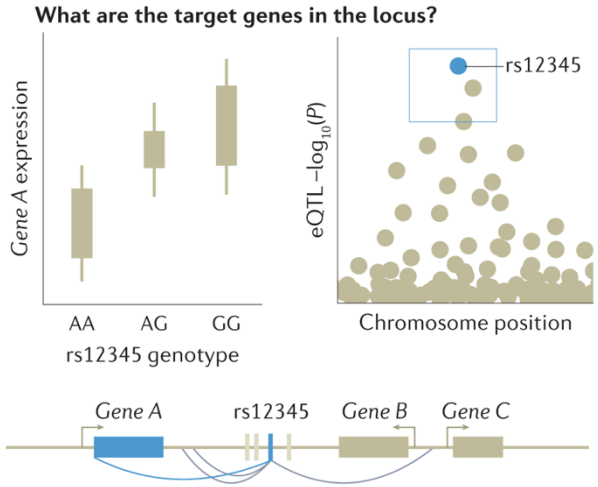
The Molecular Signatures Database (MSigDB) - resource of annotated gene sets

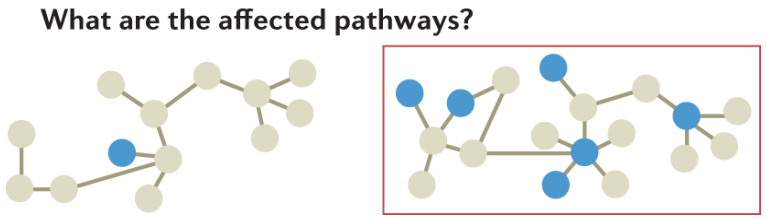
- Online pathway databases: KEGG, Biocarta, Reactome, WikiPathwyas
- Biomedical literature
- Contributed by individual domain experts

Gene Ontology (GO) Consortium - source of information on the functions of genes

Recap: SNP → Gene → Pathway







Uffelmann, E., *et al.* Genome-wide association studies. *Nat Rev Methods Primers* (2021)

Phenome-wide association study (PheWAS)

Is the variant of interest (or genetic liability) associated with other traits?

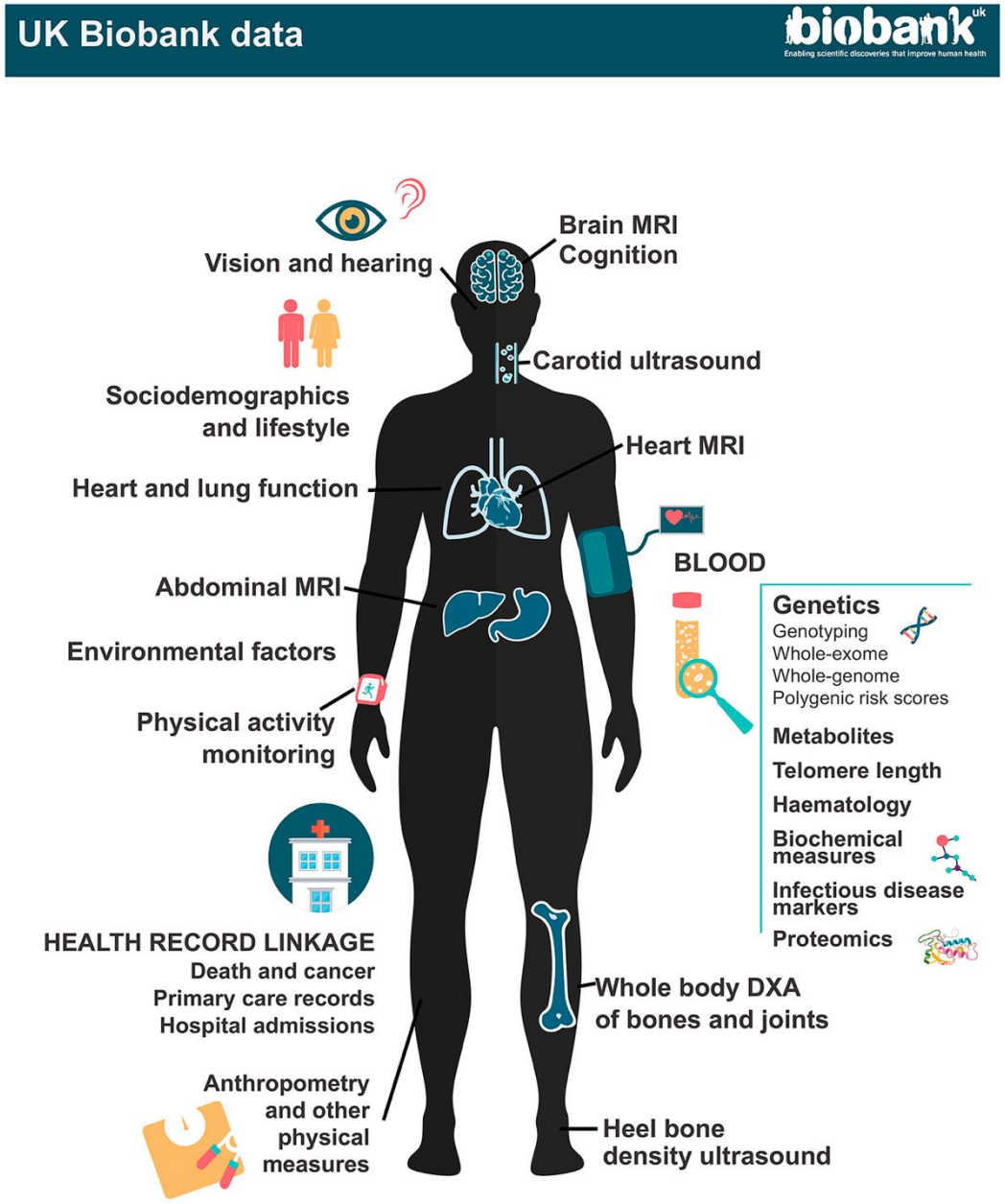
- Identify pleiotropic or mediating phenotypes or biomarkers
- Can pleiotropy explain comorbidity patterns?
- Associations unknown or underappreciated traits?
- How phenotype fits within our understanding of overall health

Requires large cohort of deeply phenotyped individuals….

biobank uk
Enabling scientific discoveries that improve human health

Slide credit: Isabelle McGrath

Biobanks contain extremely rich clinical and health information



Feng Q, et al. UK biobank: Enhanced assessment of the epidemiology and long-term impact of coronavirus disease-2019. *Cambridge Prisms: Precision Medicine*. 2023

Health-related outcomes from medial records

Primary Care

- GP data: codes about diagnoses, prescriptions and referrals
- 230,000 participants up to 2016 or 2017
- Read v2 and v3 classification system

Cancer Register

- Cancer diagnoses from national cancer registries.
- ICD9 & ICD10 codes & cancer histology code

First Occurrences

- Date that a clinical code first occurred from any source
- Provided for all 3-character ICD10 codes

Hospital Inpatient

- Hospital admission data: date, diagnosis & procedures
- ICD9 & ICD10 codes

Death Register

- Date and cause of death
- ICD10 codes

Algorithmically-defined outcomes

- Generated from multiple sources for key set of conditions

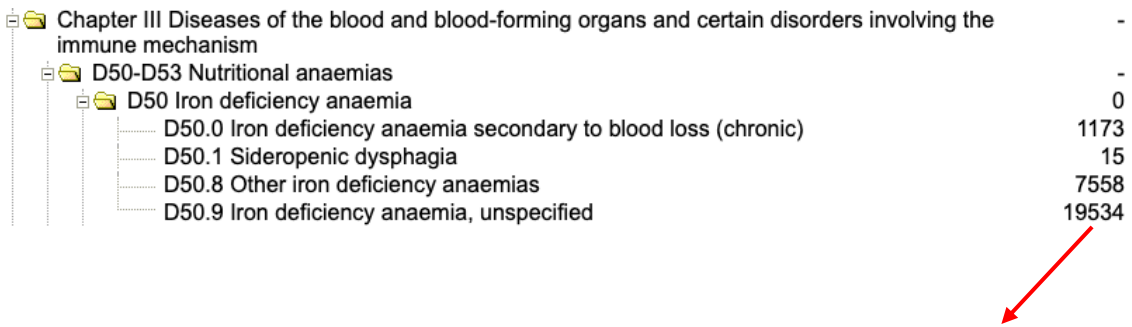
https://www.ukbiobank.ac.uk/about-our-data/types-of-data/healthcare-records/

Hospital inpatient data: ICD10 codes

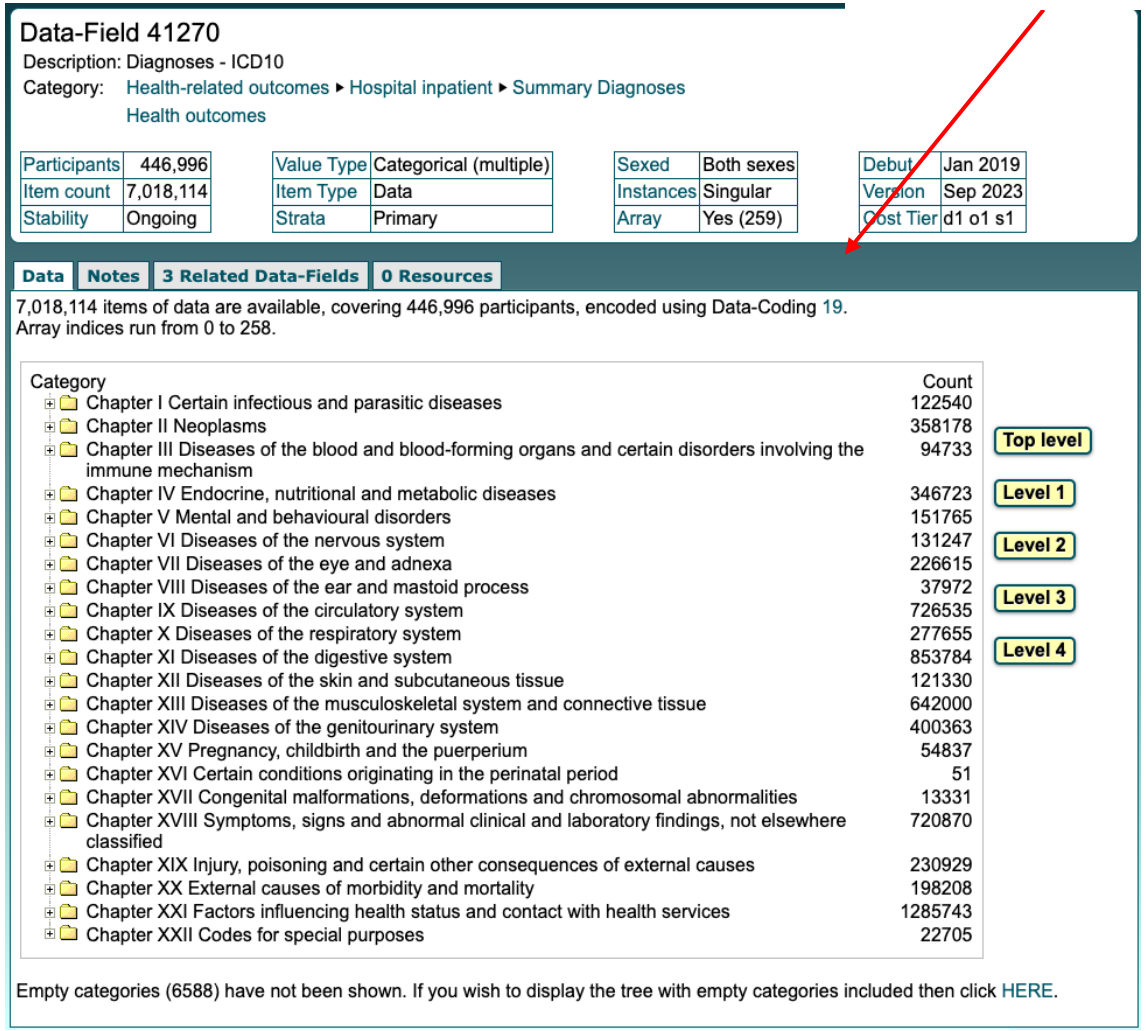
use data coding file to decrypt data

- Diagnostic codes used for billing
  - Data of diagnosis
  - Main reason for admission
  - Additional diagnoses

- Grouped into 22 chapters



number of individuals with the code

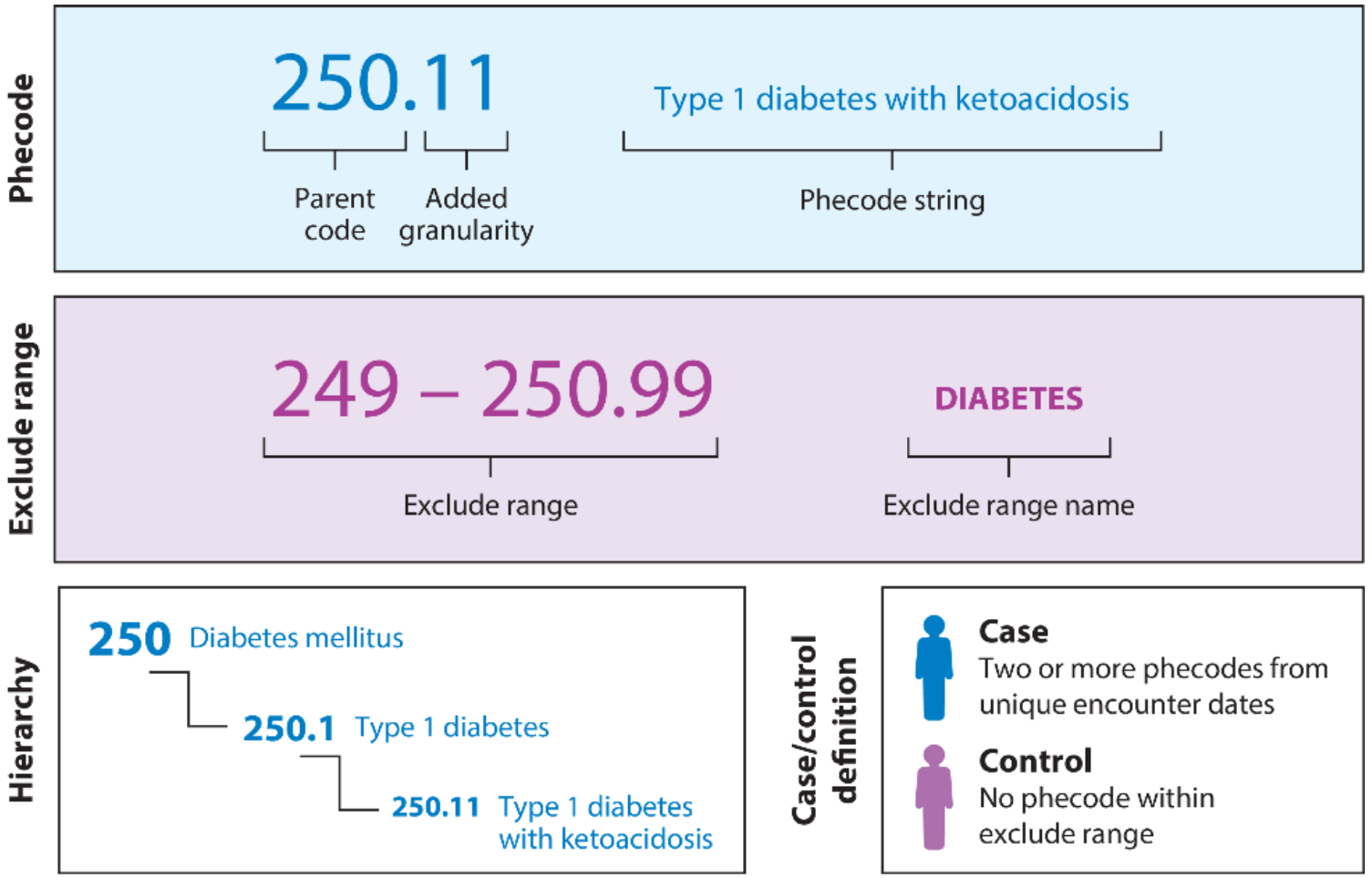


https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=41270

Phecodes are mapping groups of ICD codes for research

Lots of specific ICD10 codes → low power for discovery

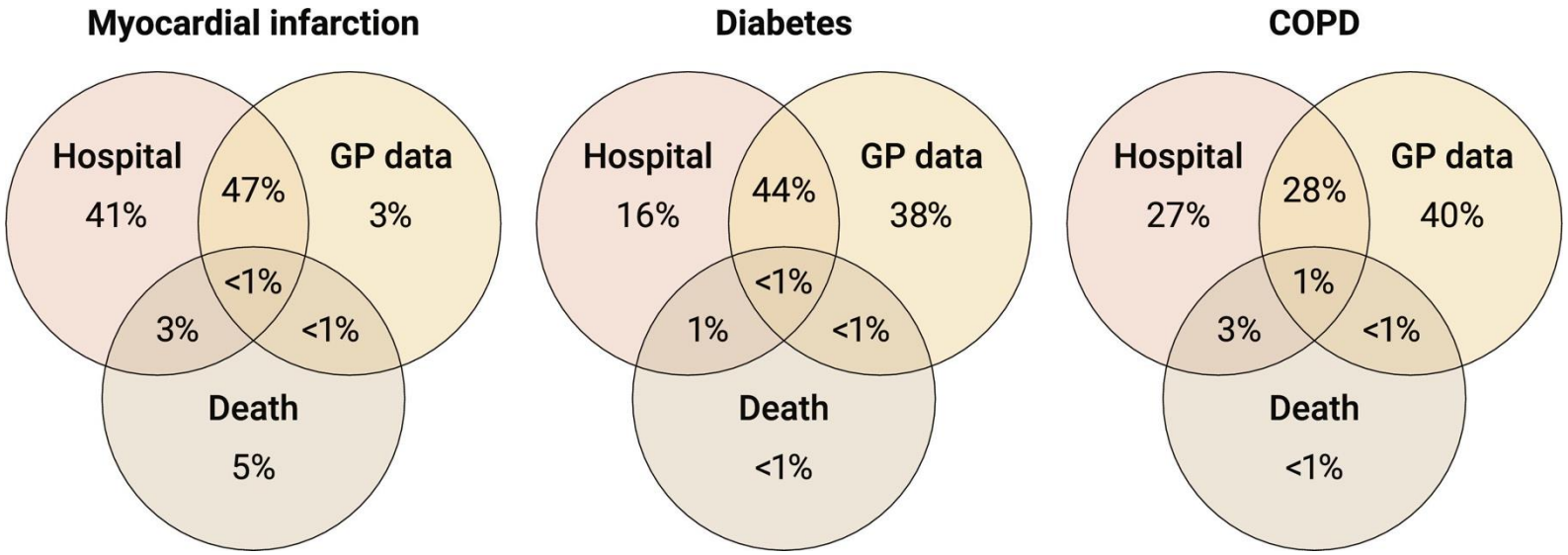
Phecodes: a high-throughput strategy for defining phenotypes using ICD10 codes



Bastarache L. 2021
Annu. Rev. Biomed. Data Sci. 4:1–19

Lisa Bastarache. 2021. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annual Review Biomedical Data Science*.

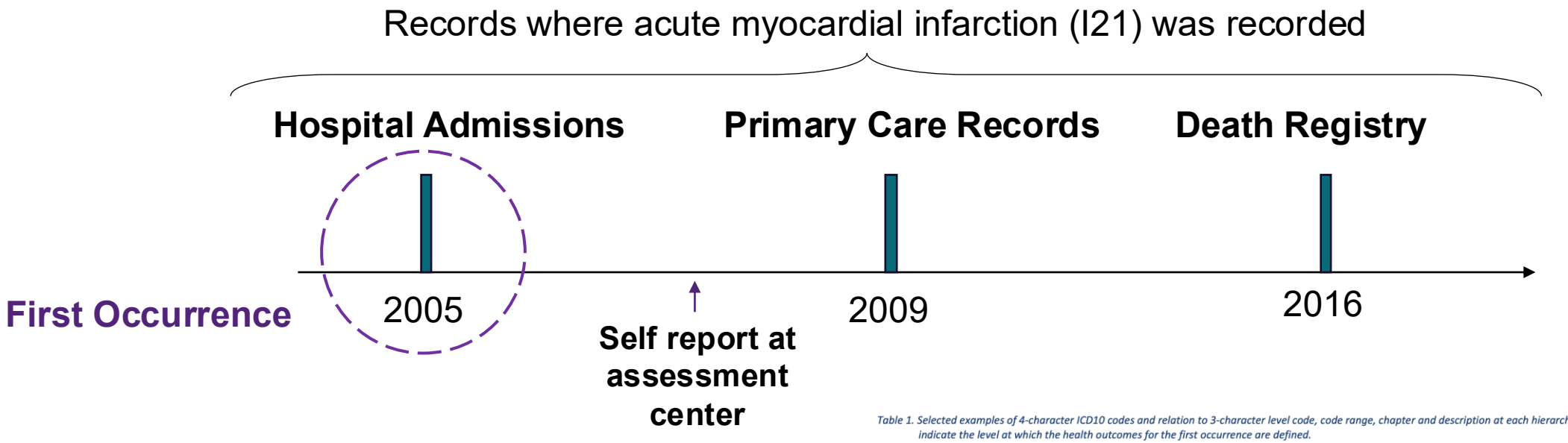
Health-related outcomes from medial records



- Identify more health outcomes by combining data sources
- Primary care (GP) includes cases identified at an earlier and less severe stage

Allen et al. Science Translation Medicine, 2024

First occurrence dataset



- Created by UKB - available for a wide range of health outcomes
- Mapped to a 3-character ICD10 code.
- Same 3-character ICD10 code ≠ Same Disease

Table 1. Selected examples of 4-character ICD10 codes and relation to 3-character level code, code range, chapter and description at each hierarchical level. The green cells indicate the level at which the health outcomes for the first occurrence are defined.

ICD10 chapter	Grouped 3-character ICD10		3-character ICD10		4-character ICD10	
	Code range	Description	Code	Description	Code	Description
Chapter IX Diseases of the circulatory system	I20-I25	Ischaemic heart diseases	I21	Acute myocardial infarction	I21	Acute myocardial infarction
					I210	Acute transmural myocardial infarction of anterior wall
					I211	Acute transmural myocardial infarction of inferior wall
					I212	Acute transmural myocardial infarction of other sites
					I213	Acute transmural myocardial infarction of unspecified site
					I214	Acute subendocardial myocardial infarction
					I219	Acute myocardial infarction, unspecified
			I22	Subsequent myocardial infarction	I22	Subsequent myocardial infarction
					I220	Subsequent myocardial infarction of anterior wall
					I228	Subsequent myocardial infarction of other sites
					I229	Subsequent myocardial infarction of unspecified site
					I221	Subsequent myocardial infarction of inferior wall
	I26-I28	Pulmonary heart disease and diseases of pulmonary circulation	I26	Pulmonary embolism	I26	Pulmonary embolism
					I260	Pulmonary embolism with mention of acute cor pulmonale
					I269	Pulmonary embolism without mention of acute cor pulmonale

Algorithmically-defined outcomes

Generated from algorithmic combinations of coded information:

- Self-reported medical condition codes reported at the baseline assessment
- Hospital admissions; diagnoses and procedures
- ICD-10 codes in Death Register records

8 Sub-Categories | 1 Parent Category | 3 Resources

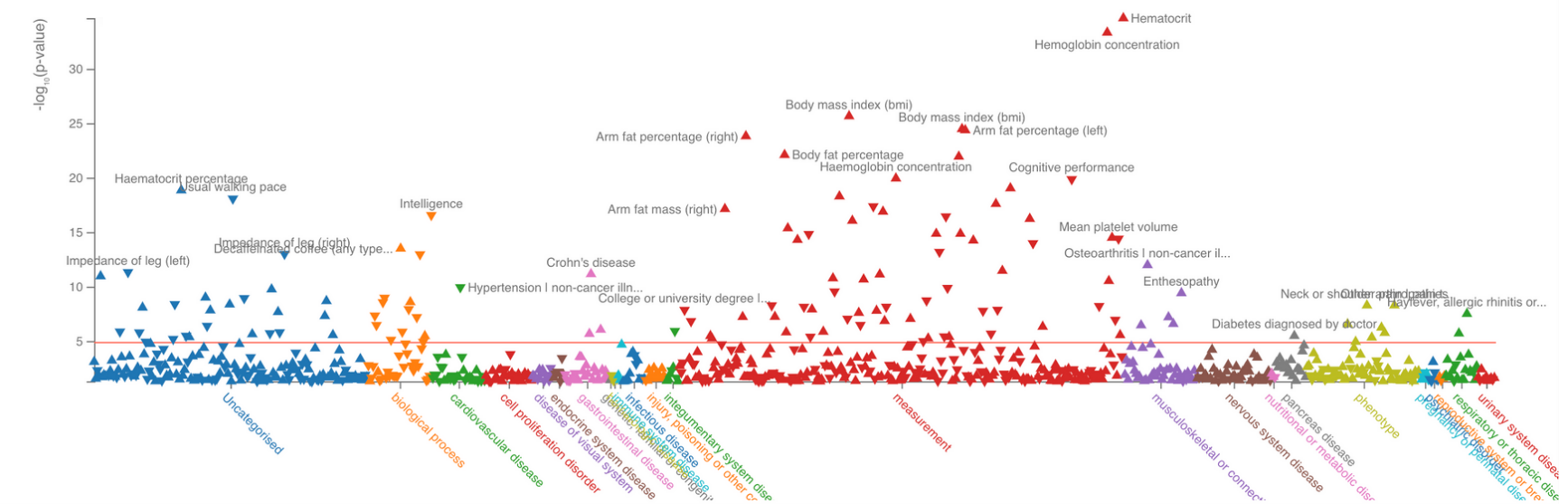
Category ID	Description	Items
45	Asthma outcomes	2
46	COPD outcomes	2
47	Dementia outcomes	8
48	End stage renal disease outcomes	2
49	Motor neurone disease outcomes	2
44	Myocardial infarction outcomes	6
50	Parkinson's disease outcomes	8
43	Stroke outcomes	8

https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=42

Readily Available PheWAS Results

OpenTargets Genetics

Associations from UK Biobank, FinnGen and/or GWAS Catalog available

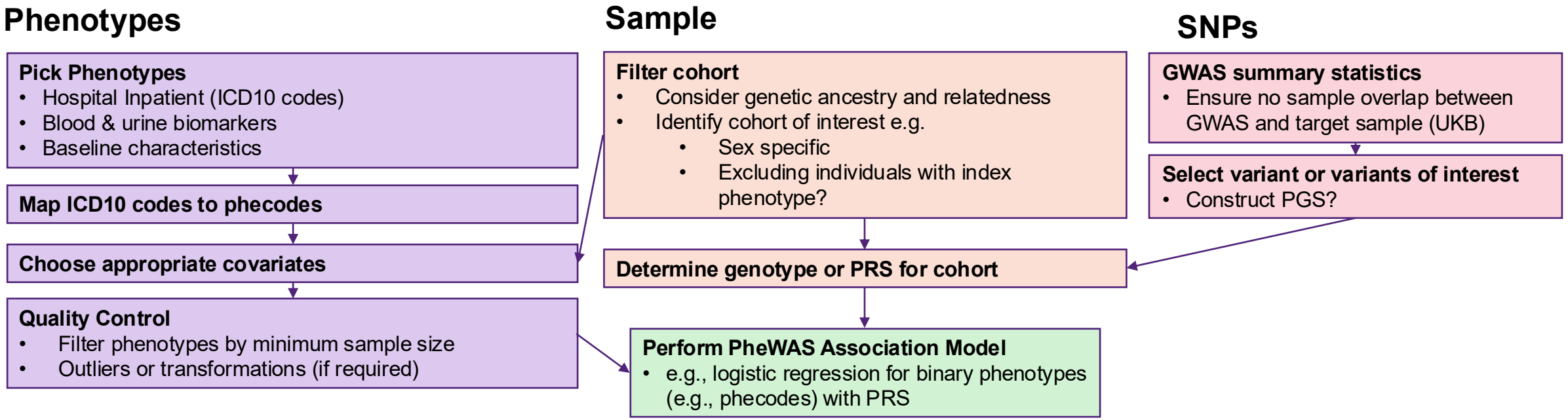


Only traits with P-value < 0.005 are shown

https://genetics.opentargets.org/

Performing your own PheWAS

- Use a PGS to investigate pleiotropic effects of genetic liability
- Interested in a specific cohort of individuals e.g. sex, specific comorbidity
- Compare effects individuals with/without the condition



- *For each Phecode*:

  glm(Phecode1 ~ PRS + covariates, data = dataframe, family = "binomial")

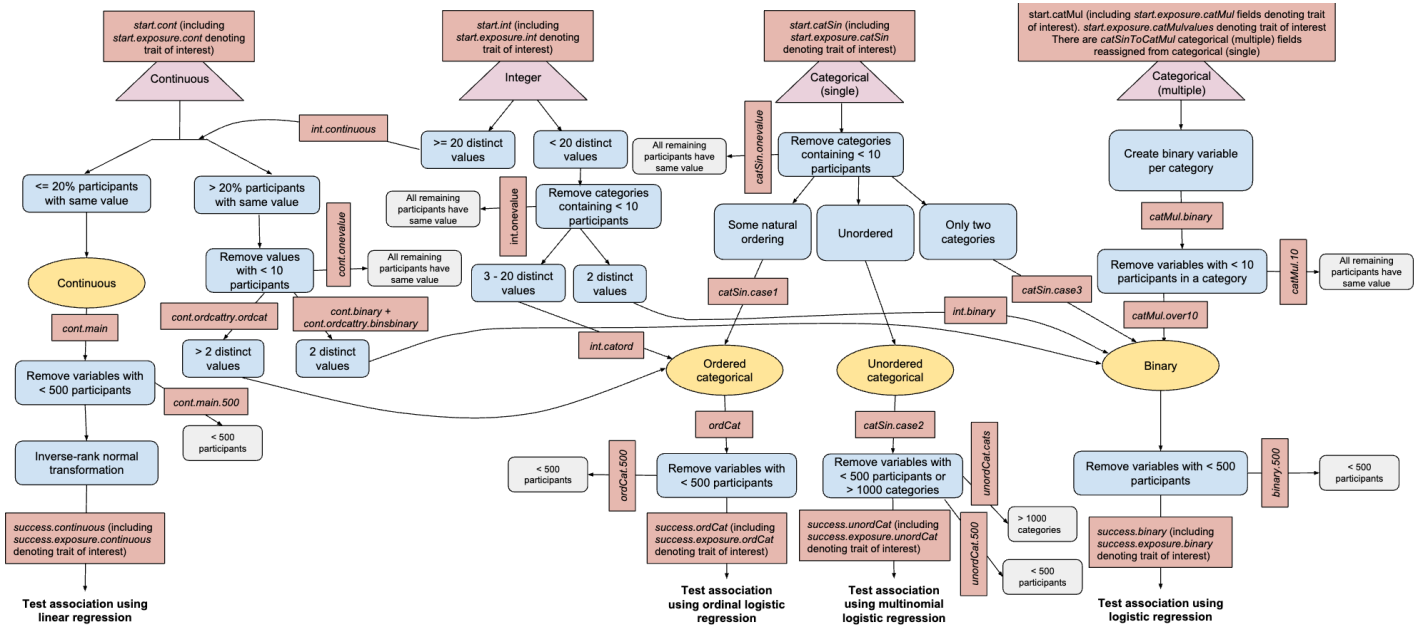
- *Multiple testing correction*

Tools available for PheWAS

Int J Epidemiol. 2018 Feb; 47(1): 29–35.
Published online 2017 Oct 5. doi: 10.1093/ije/dyx204
PMCID: PMC5837456
PMID: 29040602

Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank

Reviewed by Louise AC Millard,[1,2] Neil M Davies,[1] Tom R Gaunt,[1] George Davey Smith,[1] and Kate Tilling[1]



JOURNAL ARTICLE

pyPheWAS Explorer: a visualization tool for exploratory analysis of phenome-disease associations

Cailey I Kerley, Tin Q Nguyen, Karthik Ramadass, Laurie E Cutting, Bennett A Landman, Matthew Berger

JAMIA Open, Volume 6, Issue 1, April 2023, ooad018,
https://doi.org/10.1093/jamiaopen/ooad018
Published: 03 April 2023 **Article history**

JOURNAL ARTICLE

R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment

Robert J. Carroll, Lisa Bastarache, Joshua C. Denny Author Notes

Bioinformatics, Volume 30, Issue 16, August 2014, Pages 2375–2376,
https://doi.org/10.1093/bioinformatics/btu197
Published: 14 April 2014 **Article history**

Analyses downstream of GWAS

- Goal is to move beyond association mapping to mechanism to medicine
- Introduced some of the steps to help get us there
  - **Meta analysis**
  - **SNP heritability**
  - **Conditional analysis**
  - **Fine-map**
  - **Gene based tests**
  - **Gene-set/Pathway enrichment**
  - **PheWAS**

More tomorrow..

- **Polygenic Score Methods**
- **Mendelian Randomization**