GWAS Meta-analysis

Instructor: Daniel Howrigan
howrigan@broadinstitute.org
Session assistance: Daniel Gustavson
Daniel.Gustavson@colorado.edu

2025 International Statistical Genetics Workshop

Slides adapted from various presentations (S Medland, R Walters, W Zhou)

meta

Adjective

“Referring to itself or to the conventions of its genre”

“Self-referential”

e.g. “An analysis of the available analyses”


COURTESY OF FACEBOOK/MARK ZUCKERBERG
MARK ZUCKERBERG
META CEO


COURTESY OF FACEBOOK/MARK ZUCKERBERG
MARK ZUCKERBERG
META CEO

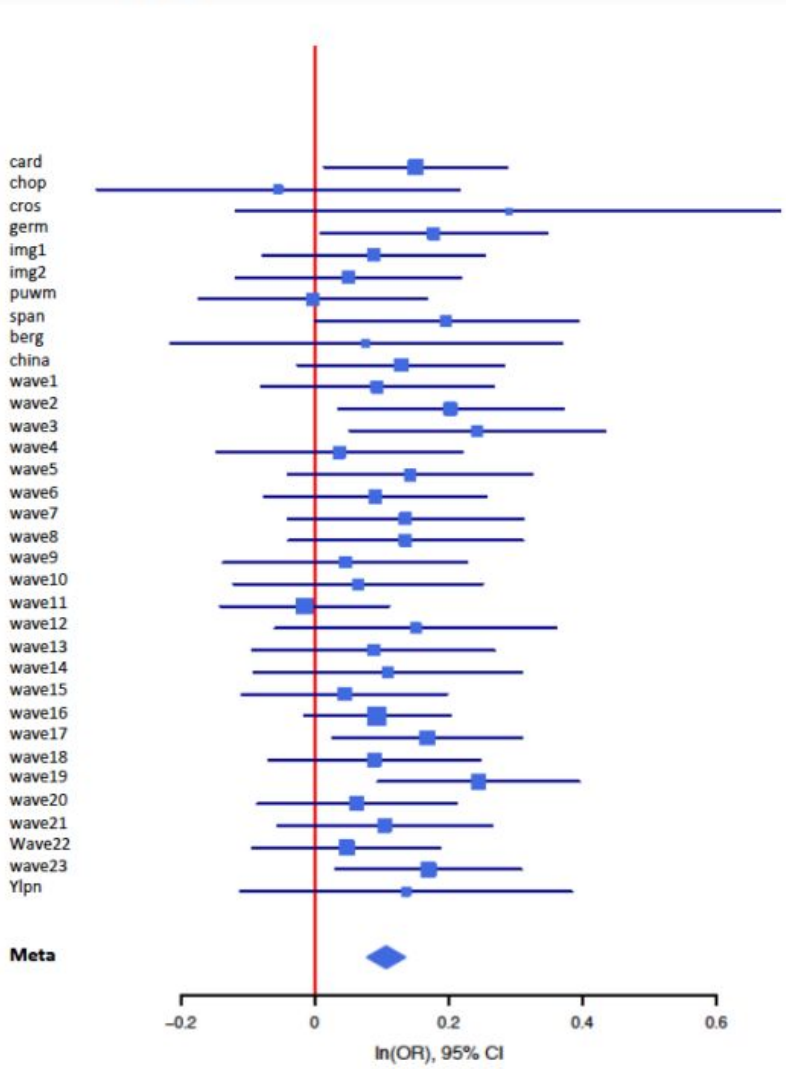
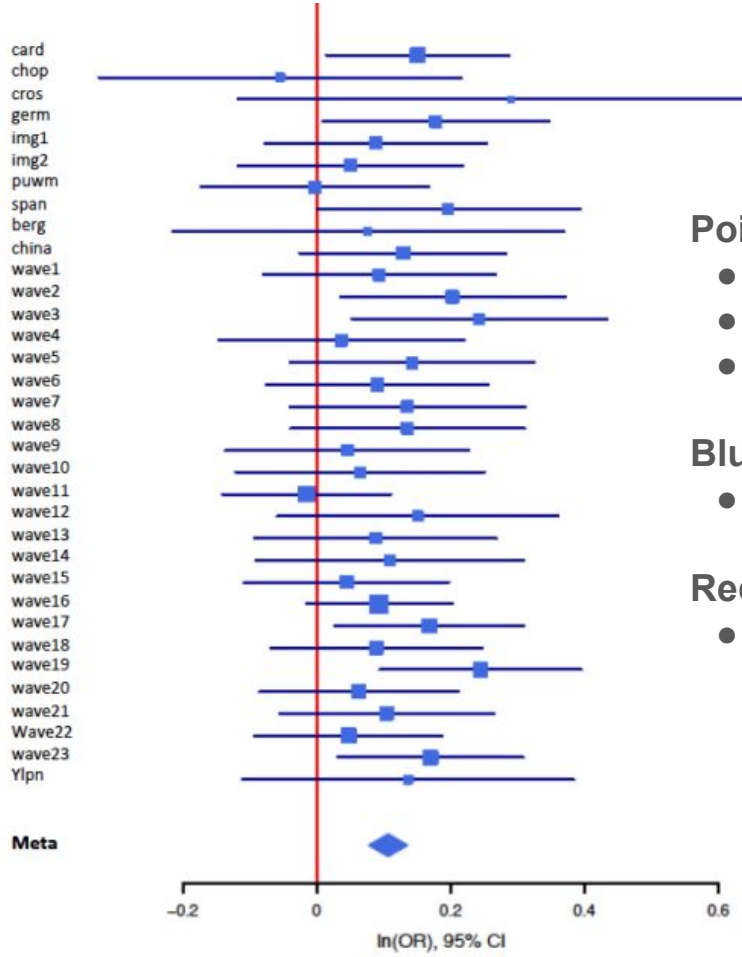


Figure from [ADHD PGC meta-analysis]

Single SNP Meta-analysis forest plot

Single cohort GWAS

Combined meta-analysis



Point (or effect) estimate

- Log Odds (Case-control)
- Beta (quantitative trait)
- Box size = sample size (N)

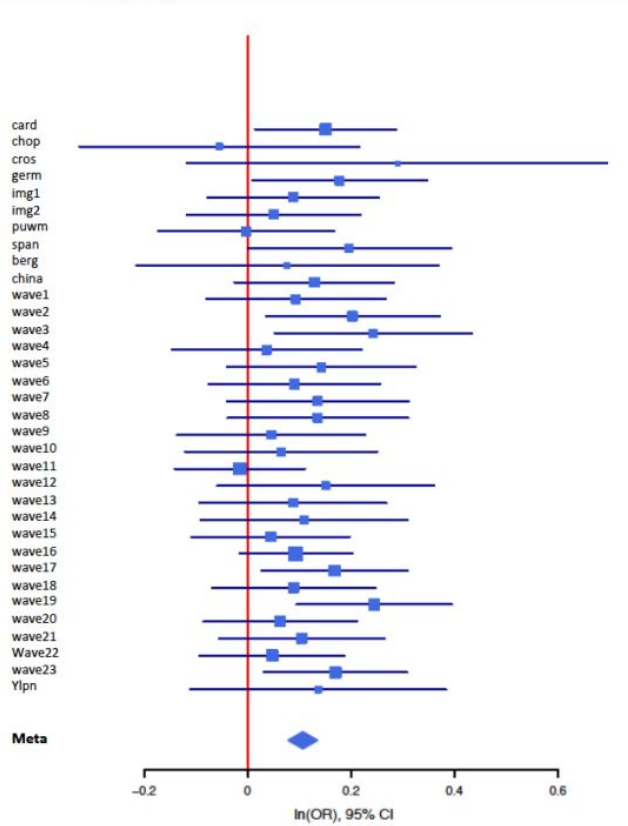
Blue line: 95% Confidence Interval

- 1.96*Std.Error

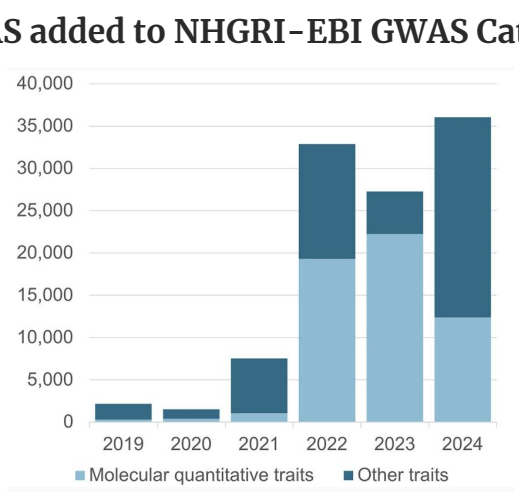
Red line

- Null hypothesis (no effect)

Benefits of meta-analysis

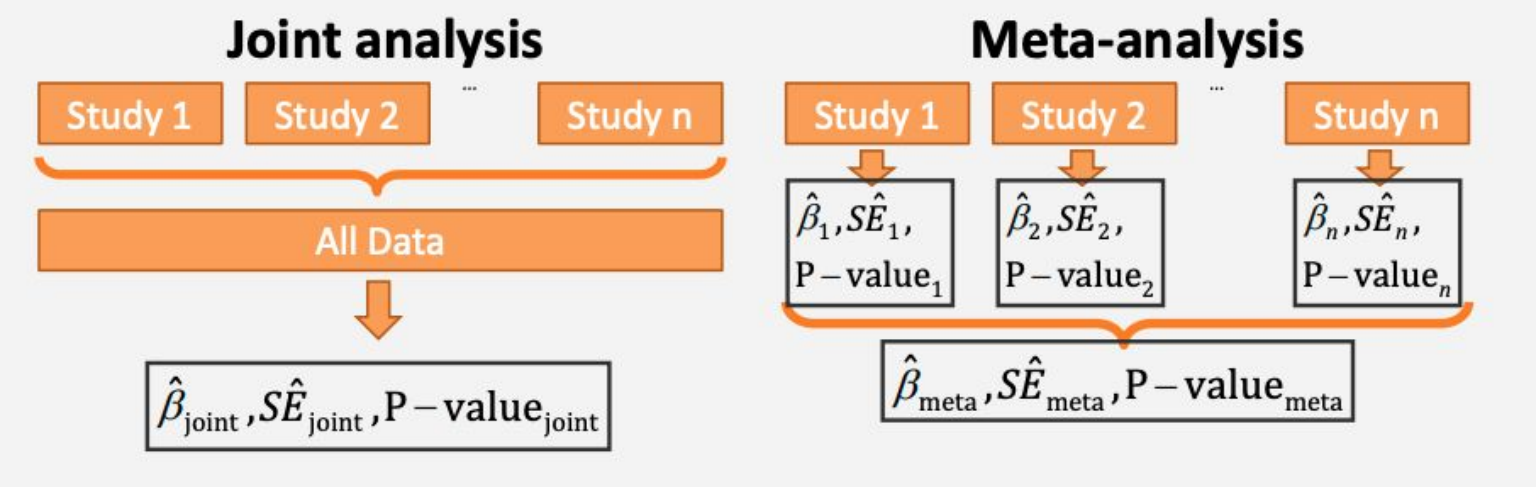


- More POWER!
- Leverage the portability of summary statistics
- Explore / expose cohort-level heterogeneity
- Replicate findings



GWAS catalog paper

Joint (“Mega”) analysis vs Meta analysis



- Common SNPs have similar power in either approach
- Meta-analysis model can handle cohort-specific covariates better
- Joint analysis of small, ancestry+platform matched cohorts can be useful within a larger meta-analysis

Session outline (*Cookbook*)

- Key parameters of a meta-analysis (*Ingredients*)
  - Test statistics
  - Weights

- Models used (*Cooking method*)
  - Base assumptions
  - Multi-trait and multi-ancestry considerations

- Getting your summary stats ready (*Instructions*)

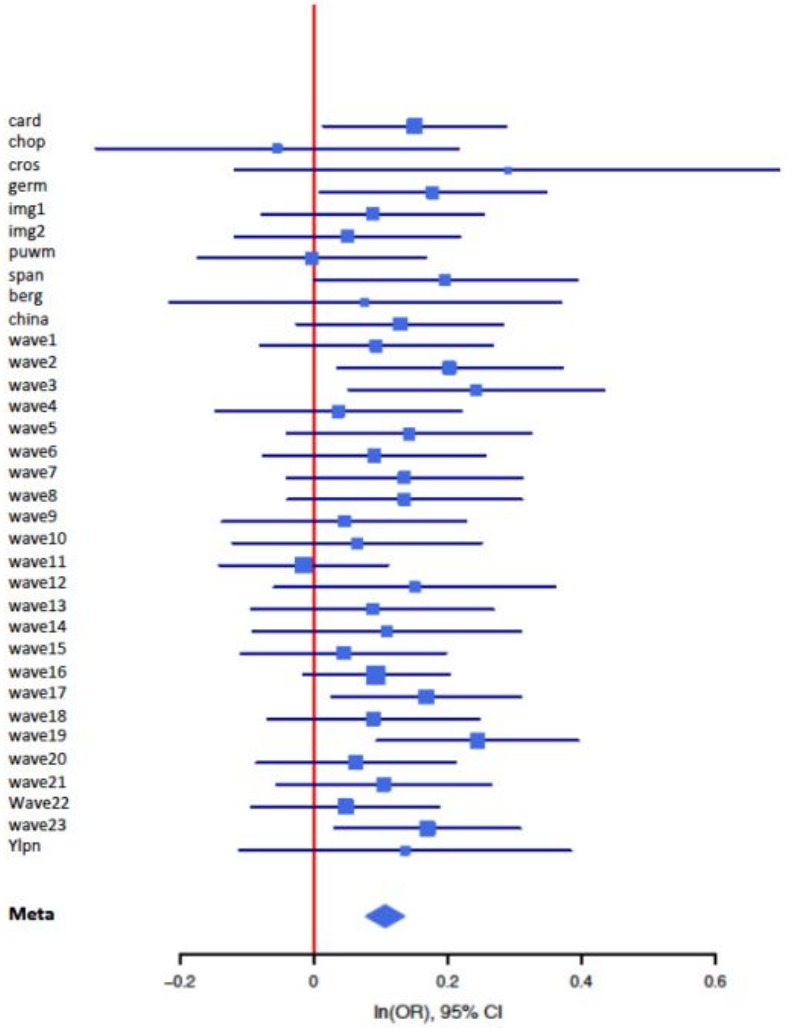
Key parameters (*Ingredients*)

Approach #1: Inverse variance weighted (IVW) method

Intuition: Give more weight to effect estimates with tighter variance when combining across all effects

Parameters used:

- **Beta** for quantitative trait
- **Log Odds Ratio / Z-score** for case-control trait
- **Standard Error** (variance estimate)



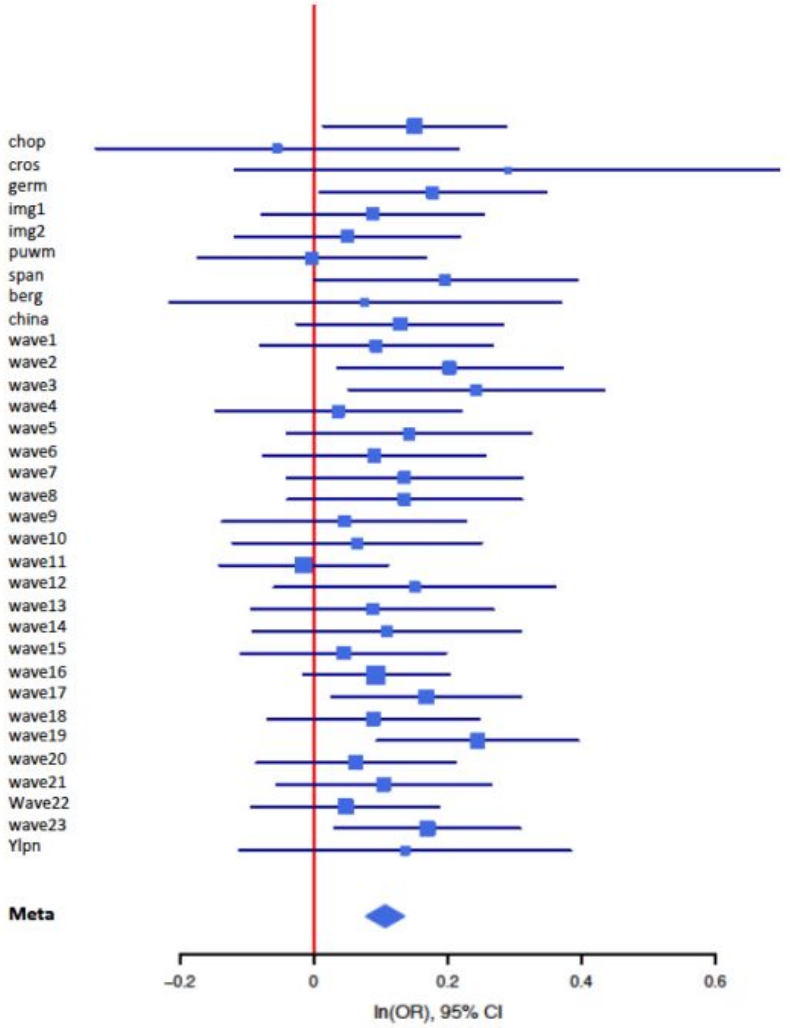
Approach #1: Inverse variance weighted (IVW) method

Intuition: Give more weight to effect estimates with tighter variance when combining across all effects

Weighted Beta example:
σ^2_i = squared standard error for the ith cohort

$$\hat{\beta} = \frac{\sum_{i=1}^{m} w_i \hat{\beta}_i}{\sum_{i=1}^{m} w_i} \qquad w_i = \frac{1}{\sigma_i^2}$$

$$SE^* = \sqrt{\frac{1}{\sum_{i=1}^{m} \frac{1}{\sigma_i^2}}}$$



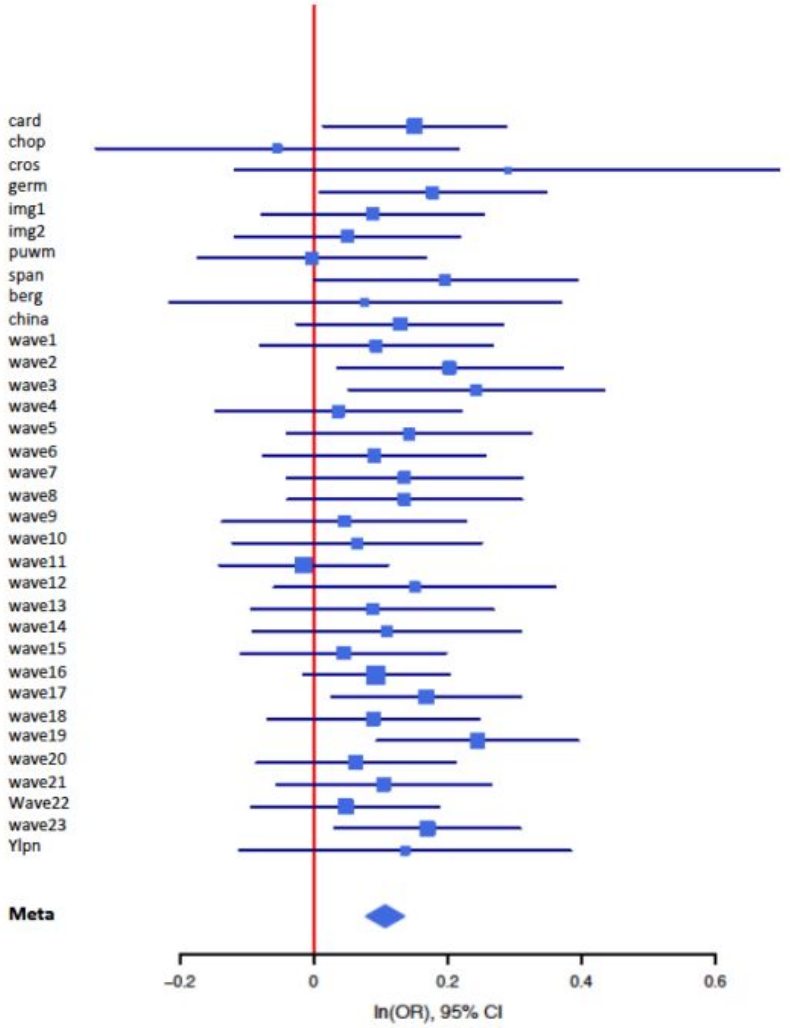
Approach #2: sample size weighted method

Intuition: Give more weight to p-values with larger sample size when combining across all effects

Parameters used:

- ***p*-values** + direction of effect (converted to Z-scores)
- **Sample sizes (n)**

$$Z = \frac{\sum_{i=1}^{m} w_i Z_i}{\sqrt{\sum_{i=1}^{m} w_i^2}} \qquad w_i = \sqrt{n_i}$$

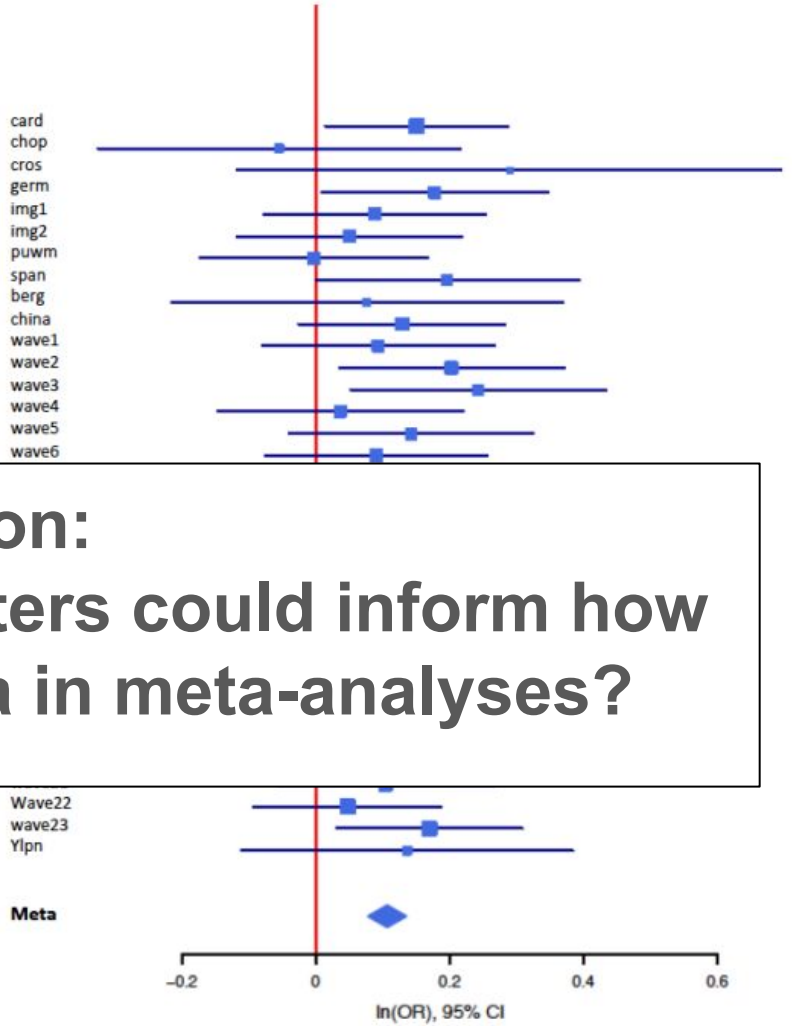


Approach #2: sample size weighted method

Intuition: Give more weight to p-values with larger sample size when combining across all effects

Parameters used:

- ***p*-values** + direction of effect (converted to Z-scores)
- **Sample sizes (n)**



Question:

What other study parameters could inform how you weight/include data in meta-analyses?

Other meta-analytical methods (rarely used in GWAMA)

- Fisher's method:

$$T = -2\sum_{i=1}^{m} \ln(p_i) \sim \chi^2_{2m}$$

- Sum of χ^2's

$$T = \sum_{i=1}^{m} Z_i^2 \sim \chi^2_m$$

These do not account for direction of effect

Session outline (*Cookbook*)

- Key parameters of a meta-analysis (*Ingredients*)
  - Test statistics
  - Weights

- **Models used (*Cooking method*)**
  - **Base assumptions**
  - **Multi-trait and multi-ancestry considerations**

- Getting your summary stats ready (*Instructions*)

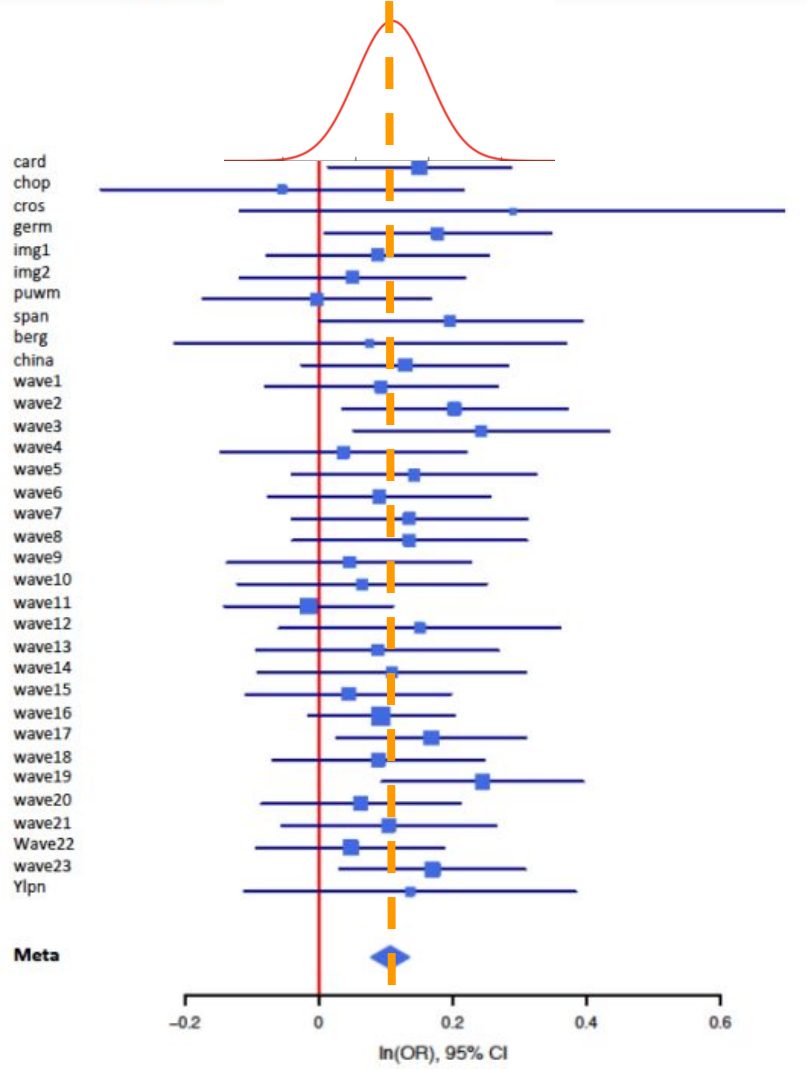
Fixed effect vs Random effects model

Fixed effect model

- Assumes the SNP has single “true” effect on the trait across all cohorts
- Error is assumed to only be “within” studies

PRO: More powerful than random effects in general

CON: Sensitive to errors in trait scaling, phenotype heterogeneity



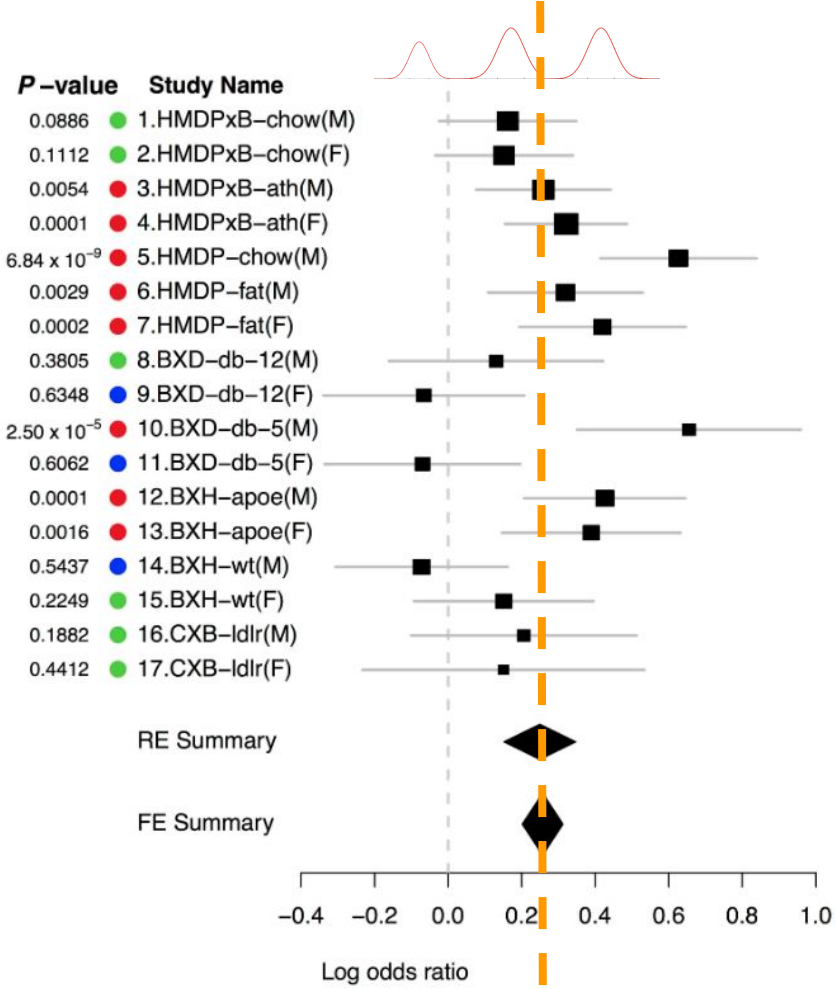
Fixed effects vs Random effects model

Random effects model

- Assumes the SNP effect on the trait varies between cohorts
- Error is assumed to be both “within” and “between” studies

PRO: Robust in the presence of effect size heterogeneity

CON: Underpowered relative to fixed effects



http://genetics.cs.ucla.edu/meta/

Fixed effects vs Random effects model

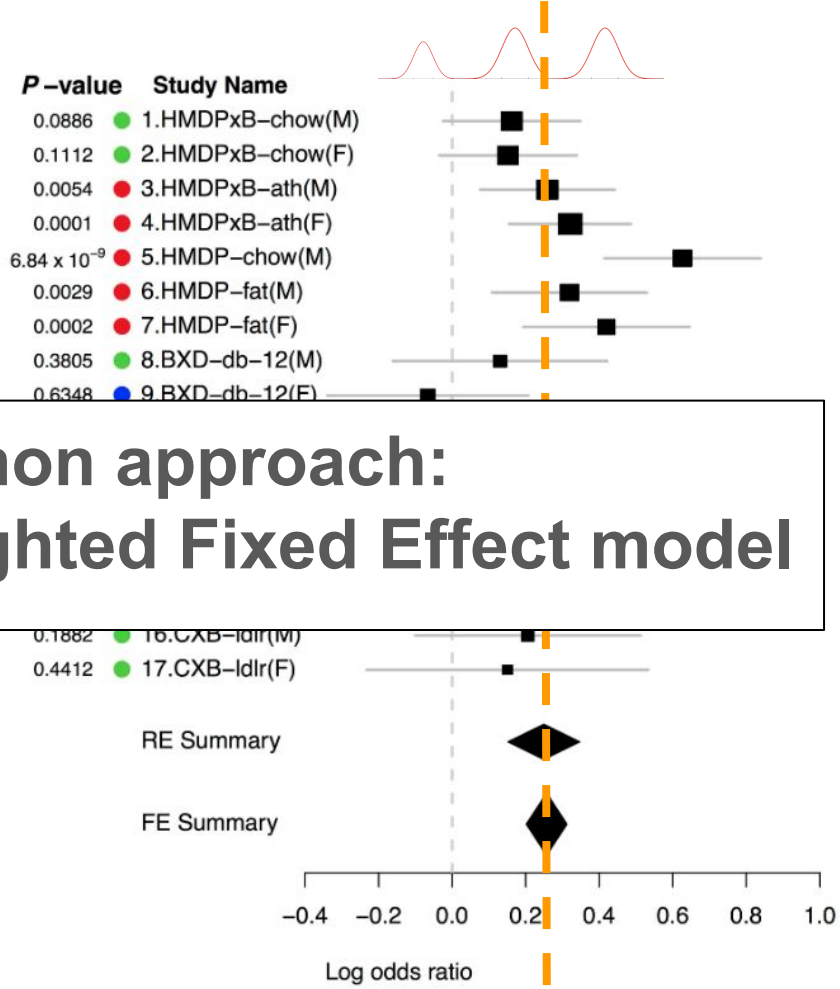
Random effects model

-
-

Most common approach:
Inverse Variance Weighted Fixed Effect model

PRO: Robust in the presence of effect size heterogeneity

CON: Underpowered relative to fixed effects



http://genetics.cs.ucla.edu/meta/

Metasoft: Why not look at both (and more)?

METASOFT provides the following methods:

Fixed Effects model (FE)

Fixed effects model based on inverse-variance-weighted effect size.

Random Effects model (RE)

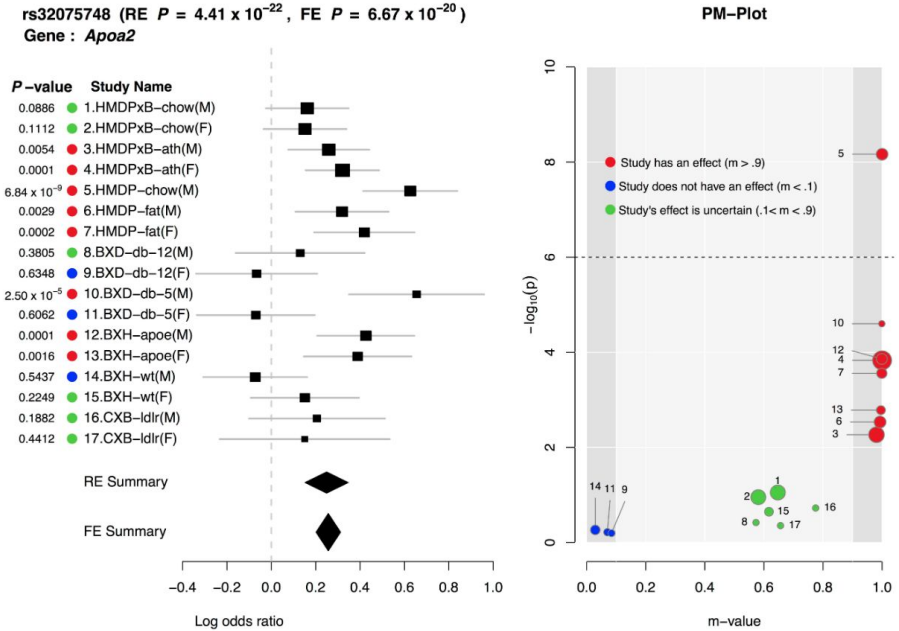
Conventional random effects model based on inverse-variance-weighted effect size (very conservative).

Han and Eskin's Random Effects model (RE2)

New random effects model optimized to detect associations under heterogeneity. (Han and Eskin, AJHG 2011)

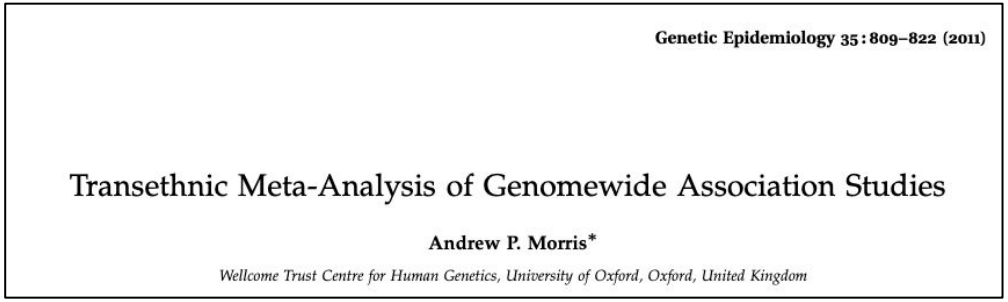
Binary Effects model (BE)

New random effects model optimized to detect associations when some studies have an effect and some studies do not. (Han and Eskin, PLoS Genetics 2012)



http://genetics.cs.ucla.edu/meta/

Bayesian models



Genetic Epidemiology 35:809–822 (2011)

Transethnic Meta-Analysis of Genomewide Association Studies

Andrew P. Morris*

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

MANTRA: Uses a Bayesian partition model to heterogeneity among clustered ancestry groups

PLOS Computational Biology

Browse Publish

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

SMetABF: A rapid algorithm for Bayesian GWAS meta-analysis with a large number of studies included

Jianle Sun, Ruiqi Lyu, Luojia Deng, Qianwen Li, Yang Zhao, Yue Zhang

Version 2 Published: March 14, 2022 • https://doi.org/10.1371/journal.pcbi.1009948

SMetABF: Asymptotic Bayes Factor approach with shotgun stochastic search (SSS) to improve the Bayesian GWAS meta-analysis framework

Using meta-regression (MR-MEGA) to model multi-ancestry heterogeneity

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

ASSOCIATION STUDIES ARTICLE

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

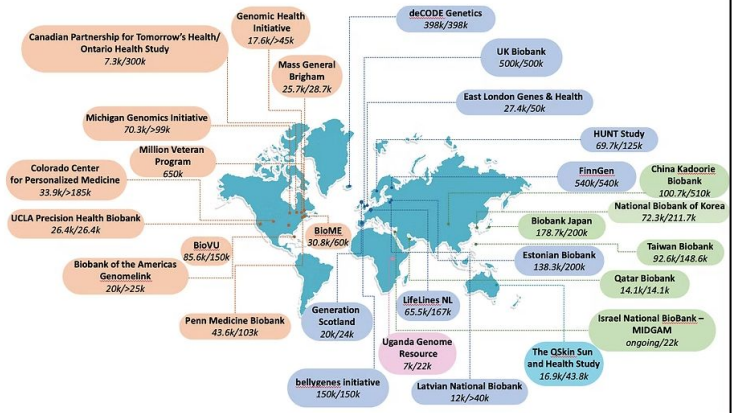
Reedik Mägi[1], Momoko Horikoshi[2,3], Tamar Sofer[4], Anubha Mahajan[2], Hidetoshi Kitajima[2], Nora Franceschini[5], Mark I. McCarthy[2,6,7], COGENT-Kidney Consortium, T2D-GENES Consortium and Andrew P. Morris[1,2,8,9,*]

"We additionally used the meta regression approach implemented in MR-MEGA (Mägi et al., 2017) to conduct the all-biobank meta analysis across all ancestries. In contrast with a fixed-effects, inverse variance-based meta-analysis, MR-MEGA accounts for the effect size heterogeneity across data sets."

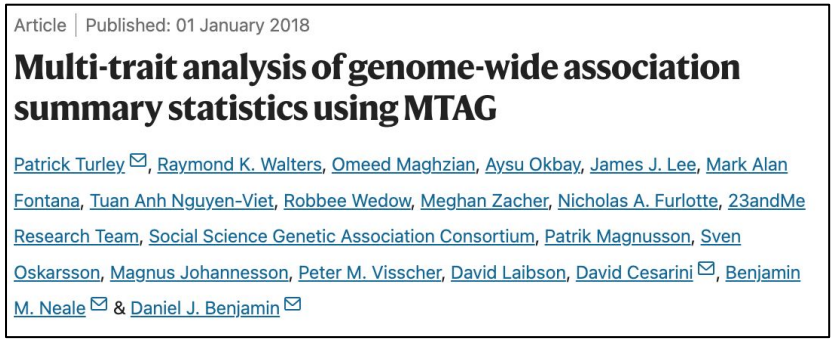
https://www.globalbiobankmeta.org/

Global Biobank Meta-analysis Initiative

29 Biobanks & Initiatives with different origins & ancestries have joined GBMI



Multi-trait meta analysis methods



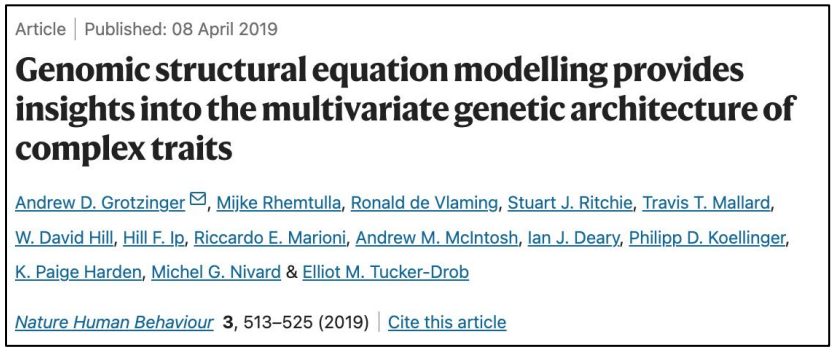
Article | Published: 01 January 2018

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 ✉

MTAG paper and Github repo

- Key assumption: all SNPs share the same variance–covariance matrix of effect sizes across traits
- Uses bivariate linkage disequilibrium (LD) score regression to account for (possibly unknown) sample overlap between the GWAS results.
- Generates trait-specific effect estimates for each SNP
- Computationally quick because every step has a closed-form solution
- Principles applied to multi-ancestry meta analysis (MAMA preprint)



Article | Published: 08 April 2019

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

Genomic SEM paper and Github repo

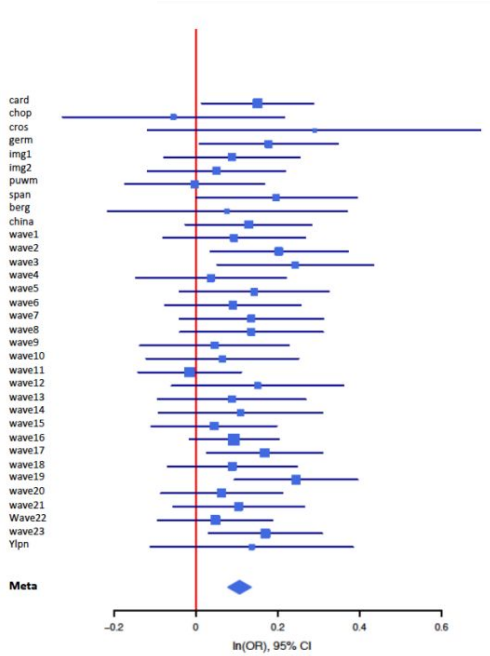
- Synthesizes genetic correlations and SNP heritabilities from GWAS summary statistics of individual traits from samples with varying and unknown degrees of overlap
- Models multivariate genetic associations among phenotypes
- Identifies variants with effects on general dimensions of cross-trait liability
- Calculates more predictive polygenic scores
- Identify loci that cause divergence between traits

Session outline (*Cookbook*)

- Key parameters of a meta-analysis (*Ingredients*)
  - Test statistics
  - Weights
- Models used (*Cooking method*)
  - Base assumptions
  - Multi-trait and multi-ancestry considerations
- Getting your summary stats ready (*Instructions*)

Running a GWAS meta-analysis

- SNP harmonization
  - SNP alignment / strand-flipping
  - Imputation reference
  - INFO score and MAF / MAC thresholds
- Sample harmonization
  - Consistency of measurement / diagnostic criteria
  - Accounting for cryptic relatedness / sample overlap
- Association model considerations
  - Model consistency across studies
  - Lambda / QC evaluation
  - Required covariates + study-specific covariates
- Interpreting meta-analysis outputs
  - Heterogeneity tests
  - Replication / leveraging external datasets



Each dataset has its own story (some longer than others..)

SNP harmonization

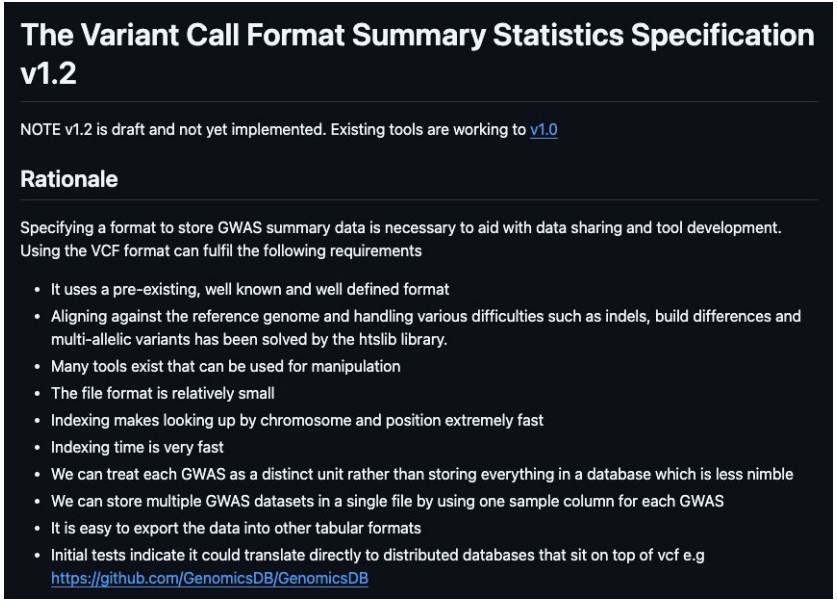
GOAL: keep as many high quality and informative SNPs as possible!

Ways to keep a lot of SNPs

- Use the same imputation reference panel across all studies
- Use alignment tools to update / format summary stats easily
  - GWAS-VCF-specification and Score GitHub repos
  - MungeSumstats in R

Reasons to drop a SNP in a specific cohort

- Strand ambiguous / palindromic SNP with high allele frequency
- Large MAF difference with ancestry-matched reference panel
- Not enough minor alleles to get informative test statistic
  - Cohort minor allele count < 20 or minimum MAF cutoff
- Low INFO / R^2 from imputation output



The Variant Call Format Summary Statistics Specification v1.2

NOTE v1.2 is draft and not yet implemented. Existing tools are working to v1.0

Rationale

Specifying a format to store GWAS summary data is necessary to aid with data sharing and tool development. Using the VCF format can fulfil the following requirements

- It uses a pre-existing, well known and well defined format
- Aligning against the reference genome and handling various difficulties such as indels, build differences and multi-allelic variants has been solved by the htslib library.
- Many tools exist that can be used for manipulation
- The file format is relatively small
- Indexing makes looking up by chromosome and position extremely fast
- Indexing time is very fast
- We can treat each GWAS as a distinct unit rather than storing everything in a database which is less nimble
- We can store multiple GWAS datasets in a single file by using one sample column for each GWAS
- It is easy to export the data into other tabular formats
- Initial tests indicate it could translate directly to distributed databases that sit on top of vcf e.g https://github.com/GenomicsDB/GenomicsDB

https://github.com/MRCIEU/gwas-vcf-specification

Sample harmonization

GOAL: understanding the trait and samples we are meta-analyzing

Trait measurement

- Is every study using the same measurement?
- How do trait means / prevalence differ across cohorts? How will this affect our meta-analysis design?

Cross-cohort sample relatedness / overlap

- Shared controls?
- Access to raw genotypes a plus
- Knowledge of sample sources often best approximation

Asthma diagnosis across biobanks

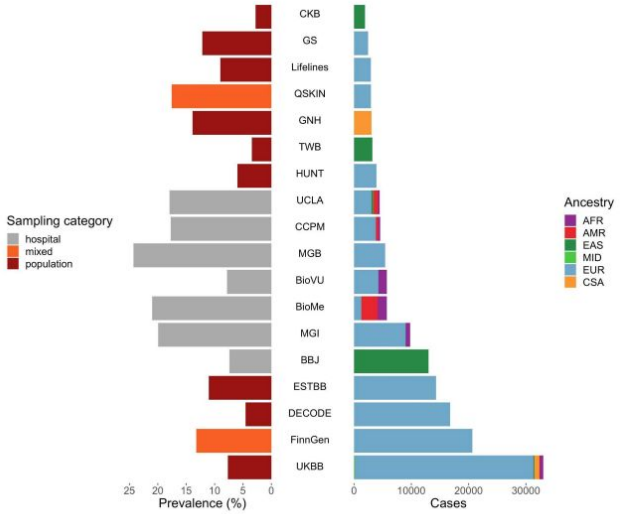


Figure 1 | 18 biobanks in GBMI contributing GWAS of asthma. Distribution of prevalence of asthma on left and number of cases of asthma on right across biobanks in GBMI. Biobanks span different sampling approaches and ancestries (AFR = African; AMR = Admixed American; EAS = East Asian; MID = Middle Eastern; EUR = European; CSA = Central and South Asian).

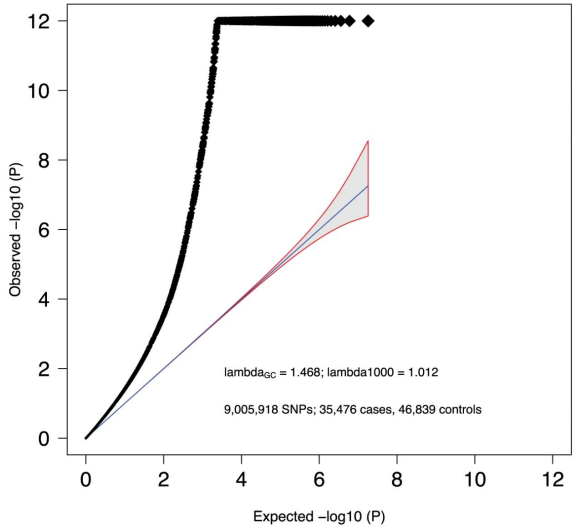
https://www.medrxiv.org/content/10.1101/2021.11.30.21267108v1.full.pdf

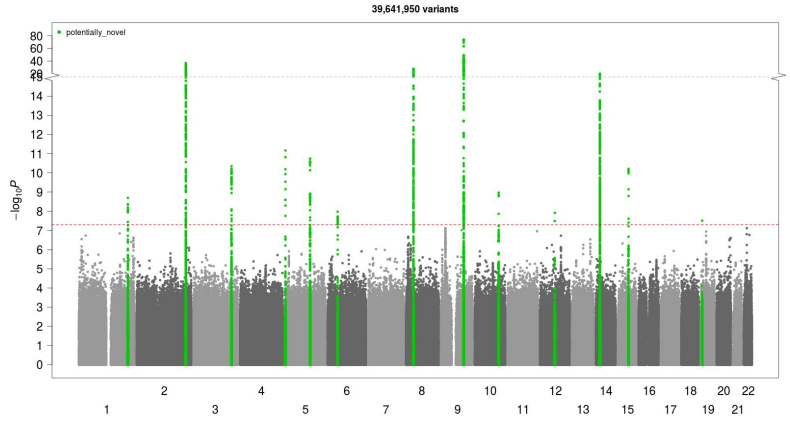
Association model considerations

GOAL: Confidence in the summary stats we are meta-analyzing

Checking your summary stats

- QQ and Manhattan plots
- Looking at lambdas / LD-score intercepts
  - Better to have clean GWAS than use GC-corrected *p*-values
- Tracking covariates across studies
  - Communicating base association model
  - Study-specific covariates needed?
- EasyQC R package





Interpreting meta-analysis outputs

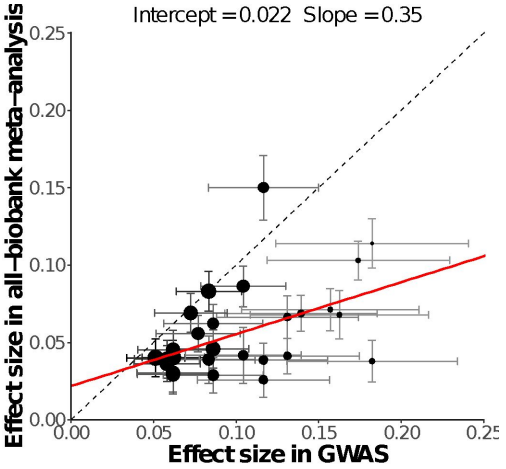
GOAL: Confidence in the meta-analysis results

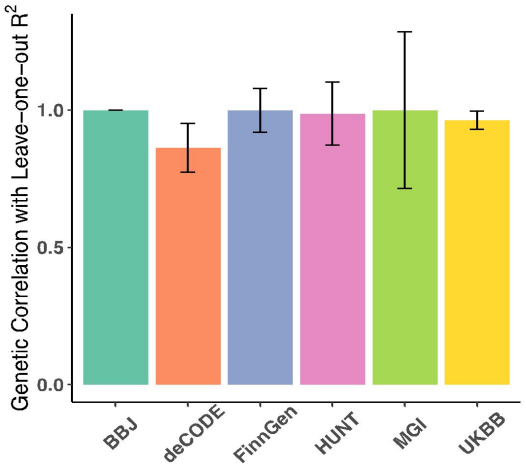
Heterogeneity tests

- Is there more variance in our effect sizes than expected?
- Cochrans Q (and *p*-value) and I^2 tests
- Usually provided in your results

QC in the "post-GWAS" era

- Comparing effect sizes with "known loci" (GWAS catalog, PheWAS scans)
- Leave-one-out analyses

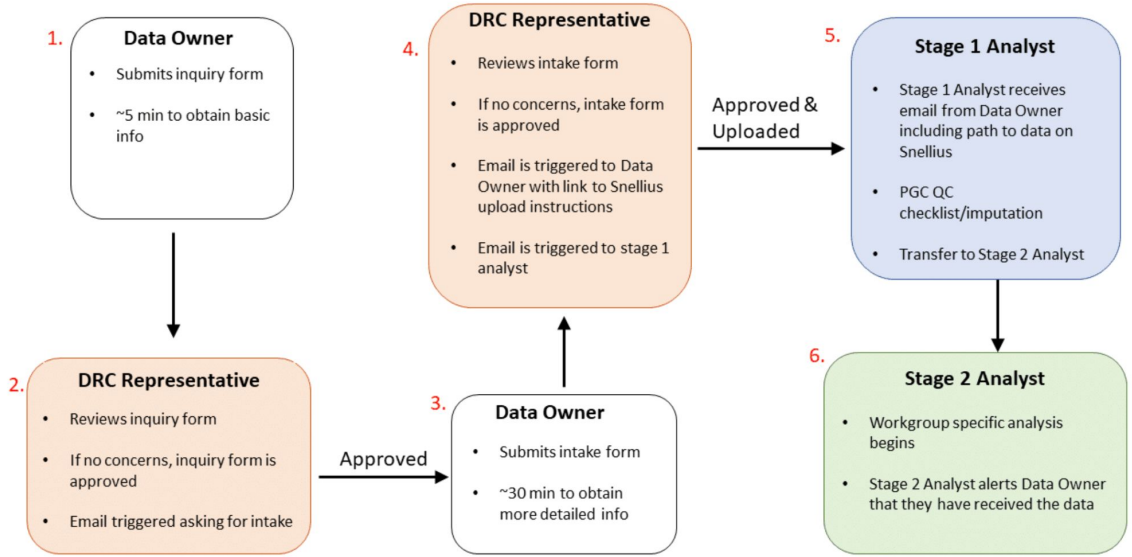




GBMI 2019 slide deck

It takes a community….

- Coordinating data sharing plans/guidelines
- Data use agreements
- When to freeze, when to unfreeze
- Sharing code vs sharing raw data
- Don't be the weakest link in the GWAMA chain
- PGC data inquiry form
  - Getting data descriptions at intake, not at paper submission



Meta analysis software

METAL

PLINK

Metasoft

GWAMA / MR-MEGA

R packages (general meta-analysis)

Meta

Metafor

Meta-analysis practical

Meta-analysis Qualtrics workbook

Workshop Directory:

```
cd /home/practicals/2.3.Meta-analysis_DanHowrigan/final
```